IJCSIS Vol. 11 No. 6, June 2013
ISSN 1947-5500

International Journal of Computer Science & Information Security


Cornell University
Library


Cogprints


Google scholar


.docstoc
find and share professional documents


ScientificCommons

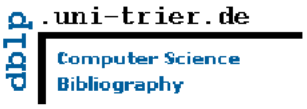

View my documents on
Scribd


BASE
Bielefeld Academic Search Engine


SCIRUS
search engine for science


SciRate.com


CiteSeerx beta


dblp .uni-trier.de
Computer Science
Bibliography


Q·Sensei


DOAJ
DIRECTORY OF
OPEN ACCESS
JOURNALS


EBSCO
HOST


ProQuest

IJCSIS

ISSN (online): 1947-5500

Please consider to contribute to and/or forward to the appropriate groups the following opportunity to submit and publish original scientific results.

CALL FOR PAPERS
International Journal of Computer Science and Information Security (IJCSIS)
January-December 2013 Issues

The topics suggested by this issue can be discussed in term of concepts, surveys, state of the art, research, standards, implementations, running experiments, applications, and industrial case studies. Authors are invited to submit complete unpublished papers, which are not under review in any other conference or journal in the following, but not limited to, topic areas.
See authors guide for manuscript preparation and submission guidelines.

Indexed by Google Scholar, DBLP, CiteSeerX, Directory for Open Access Journal (DOAJ), Bielefeld Academic Search Engine (BASE), SCIRUS, Cornell University Library, ScientificCommons, EBSCO, ProQuest and more.

Deadline: see web site
Notification: see web site
Revision: see web site
Publication: see web site

Context-aware systems
Networking technologies
Security in network, systems, and applications
Evolutionary computation
Industrial systems
Evolutionary computation
Autonomic and autonomous systems
Bio-technologies
Knowledge data systems
Mobile and distance education
Intelligent techniques, logics and systems
Knowledge processing
Information technologies
Internet and web technologies
Digital information processing
Cognitive science and knowledge
Agent-based systems
Mobility and multimedia systems
Systems performance
Networking and telecommunications
Software development and deployment
Knowledge virtualization
Systems and networks on the chip
Knowledge for global defense
Information Systems [IS]
IPv6 Today - Technology and deployment
Modeling
Software Engineering
Optimization
Complexity
Natural Language Processing
Speech Synthesis
Data Mining

For more topics, please see web site https://sites.google.com/site/ijcsis/

For more information, please visit the journal website (https://sites.google.com/site/ijcsis/)

Editorial
Message from Managing Editor

International Journal of Computer Science and Information Security *(IJCSIS – established in 2009), is an online academic journal that adheres to the highest standards of peer review and engages established and emerging scholars from anywhere in the world. The journal focuses on innovative developments, research challenges/solutions in computer science and information technologies. We are pleased to mention that we are committed to placing this journal at the forefront for the dissemination of novel and exciting research output and findings in the fields of computer science. We strongly believe that the direct communication between the editors and authors are important for the quality and well being of the IJCSIS Journal and its contents.*

IJCSIS archives all publications in major academic/scientific databases; abstracting/indexing, editorial board and other important information are available online on homepage. Indexed by the following International agencies and institutions: Google Scholar, Bielefeld Academic Search Engine (BASE), CiteSeerX, SCIRUS, Cornell's University Library EI, Scopus, DBLP, DOI, ProQuest, EBSCO. Google Scholar reported a large amount of cited papers published in IJCSIS. IJCSIS supports the Open Access policy of distribution of published manuscripts, ensuring "free availability on the public Internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of [published] articles".

IJCSIS is currently accepting quality manuscripts for upcoming issues based on original qualitative or quantitative research that explore innovative conceptual framework or substantial literature review opening new areas of inquiry and investigation in Computer science. Case studies and works of literary analysis are also welcome. IJCSIS editorial board consisting of international experts solicits your contribution to the journal with your research papers, projects, surveying works and industrial experiences. IJCSIS appreciates all the insights and advice from authors and reviewers.

We look forward to your collaboration. For further questions please do not hesitate to contact us at ***ijcsiseditor@gmail.com****.*

A complete list of journals can be found at:
http://sites.google.com/site/ijcsis/

IJCSIS Vol. 11, No. 6, June 2013 Edition

ISSN 1947-5500 © IJCSIS, USA.

Journal Indexed by (among others):

IJCSIS EDITORIAL BOARD

TABLE OF CONTENTS

dynamic virtual enterprise. This algorithm of cooperation is divided into four phases, namely: publication, search, filter and connection.

Mots-clés : Workflow, WFMS, Multi-level Architecture Cooperation, Virtual Enterprise, CoopFlow, cloud.

4. Paper 31051338: Text Summarization Using Tense Recognizing Identifier (pp. 20-24)

Rajesh Wadhvani, Devshri Roy
Computer Science Department, National Institute of Technology, Bhopal, India

Abstract —Text summarization method finds important data and select relevant sentences to form a summary from the original text data. In this paper a new approach for extracting summary from a document is proposed. The updated information about the document is obtained in two phases. In the first phase, parsing is performed on tagged sentences governed by some rules and different chunks obtained are separated. In Second phase, summary of future chunks is found based on some parameters like word frequency, sentence position etc. To extract the future sense of sentences from the complete text, modal verbs are used. The new summary produced can now meets the need of user by making selection easier and saving reading time.We generated the summaries of documents using our approach and compared the results with respective ideal summaries that showed that sentences in the generated summaries are really helpful in extracting the important data according to user.

Keywords: Knowledge Extraction; Text Summarization; Extract Summarization; Abstract Summarization; Part-ofspeech (POS) Tagging; Scoring.

5. Paper 31051339: A Game Theoretic Formulation for Strategic Sensing in Cognitive Radio Networks: Equilibrium Analysis and Fully Distributed Learning (pp. 25-32)

Sofia Bouferda, RITM-ESTC-GREENTIC, ENSEM-Hassan II University, Casablanca, Morocco
Essaid Sabir, ENSEM-GREENTIC, Hassan II University, Casablanca, Morocco
Aawatif Hayar, ENSEM-GREENTIC, Hassan II University, Casablanca, Morocco
Mounir RIFI, RITM-ESTC-GREENTIC, Hassan II University, Casablanca, Morocco

Abstract — In cognitive radio (CR) systems, the licenced bands are opened to secondary users (SU), when they are not used. To support this reuse the CR users must sense the environment, and use the spectrum when its not utilized by primary users (PU). For these reasons, the spectrum sensing is an important topic in cognitive radio networks (CRN). In this paper, we consider a distributed opportunistic access (D-OSA), in which CR users attempt to access a channel licensed to a primary network. In this context, we formulate the problem of designing the equilibrium sensing time in a distributed manner, in order to maximize the throughput of CR users while guarantying a good protection to the PU. Next, we study the Nash equilibrium of the system, we also propose a combined learning algorithm for continuous actions that is fully distributed, and allows to the CR users to learn their equilibrium payoffs and their equilibrium sensing time. Finaly we give a comparison between the proposed solution and a centralized one. The simulation results show that the system can learn the sensing time and converge to a unique Nash equilibrium, which come to prove the theoretical study. The comparison between the proposed solution and the centralized one shows an expected result, which is the higher performances of the centralized method, in terms of throughput and sensing time, but as we can see in the simulation results, the difference is slight.

Keywords— Distributed Spectrum sensing, Nash equilibrium, combined learning

6. Paper 31051341: A Neural Approach for Analysis of Lymphocytes in Detection of Rheumatoid Arthritis using Center Approximation Method (pp. 33-38)

Murugan V., Thivakaran T.K.
Department of Computer science and Engineering,Sri Venkateswara College of Engineering,Chennai, India

Abstract — Rheumatoid Arthritis is an autoimmune disease that mainly affects joints in the human body. It is a chronic disease that causes stiffness, pain, swelling and limited motion of various joints which leads to the erosion of bones at the joints. As treatments are currently available only to delay or stop the erosion and not to regain the eroded joints, it is important to detect its occurrence at the onset of aberration. Early detection of Rheumatoid Arthritis involves analysis of Lymphocytes present in human blood. Manual analysis of Lymphocytes is a long term process which needs an expert hematologist for continuous microscopic assessment of blood smear which is costly and time consuming. Digital image processing plays a vital role in the field of medical diagnosis. The proposed work aims to minimize the expenditure of the assessment of Lymphocytes with the help of digital image processing techniques. 105 samples of blood smear images containing Lymphocytes are collected from various laboratories out of which 60 images are affected and 45 images are not affected by rheumatoid arthritis. In order to reduce the noise from these images, initially they are enhanced using Weiner Filter. To locate the region of interest the enhanced images are subjected to thresholding. After thresholding image is allowed to segmentation using bounding box technique in order to separate the nucleus of the Lymphocyte from the smear image. The segmented nucleus is used to extract the required features: area, perimeter and circularity. Here the circularity value is obtained using center approximation method. These features are given to a single layer perceptron which achieved a classification accuracy of about 97.50%.

Keywords - Adaptive thresholding, Circularity of Lymphocytes.

7. Paper 31051342: The Use of Cuckoo Search in Estimating the Parameters of Software Reliability Growth Models (pp. 39-46)

Dr. Najla Akram AL-Saati, Marwa Abd-AlKareem
Software Engineering Dept, College of Computer Sciences & Mathematics, Mosul, Iraq

Abstract — This work aims to investigate the reliability of software products as an important attribute of computer programs; it helps to decide the degree of trustworthiness a program has in accomplishing its specific functions. This is done using the Software Reliability Growth Models (SRGMs) through the estimation of their parameters. The parameters are estimated in this work based on the available failure data and with the search techniques of Swarm Intelligence, namely, the Cuckoo Search (CS) due to its efficiency, effectiveness and robustness. A number of SRGMs is studied, and the results are compared to Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and extended ACO. Results show that CS outperformed both PSO and ACO in finding better parameters tested using identical datasets. It was sometimes outperformed by the extended ACO. Also in this work, the percentages of training data to testing data are investigated to show their impact on the results.

Keywords- Software Reliability; Growth Models; Parameter estimation; Swarm Intelligence; Cuckoo Search

8. Paper 31051344: Toward a New Approach for Modeling Dependability of Data Warehouse System (pp. 47-54)

Imane Hilal, Nadia Afifi, Reda Filali Hilali, Mohammed Ouzzif
RITM Lab., Computer Engineering Department, ESTC, Hassan II University, Casablanca, Morocco

Abstract — The sustainability of any Data Warehouse System (DWS) is closely correlated with user satisfaction. Therefore, analysts, designers and developers focused more on achieving all its functionality, without considering others kinds of requirement such as dependability's aspects. Moreover, these latter are often considered as properties of the system that will must be checked and corrected once the project is completed. The practice of "fix it later" can cause the obsolescence of the entire Data Warehouse System. Therefore, it requires the adoption of a methodology that will ensure the integration of aspects of dependability since the early stages of project DWS. In this paper, we first define the concepts related to dependability of DWS. Then we present our approach inspired from the MDA (Model Driven Architecture) approach to model dependability's aspects namely: availability, reliability, maintainability and security, taking into account their interaction.

Keywords-component; Data Warehouse System; Model Driven Architecture ; Dependability; Availability; Reliability, Security, Maintainability.

9. Paper 31051345: A Novel Framework using Similar to Different Learning Strategy (pp. 55-67)

Bhanu Prakash Battula, Dr. R. Satya Prasad
Acharya Nagarjuna University, Guntur, India.

Abstract — Most of the existing classification techniques concentrate on learning the datasets as a single similar unit, in spite of so many differentiating attributes and complexities involved. However, traditional classification techniques, require to analysis the dataset prior to learning and for not doing so they loss their performance in terms of accuracy and AUC. To this end, many of the machine learning problems can be very easily solved just by careful observing human learning and training nature and then mimic the same in the machine learning. In response to these issues, we present a comprehensive suite of experiments carefully designed to provide conclusive, reliable, and significant results on the problem of efficient learning. This paper proposes a novel, simple and effective machine learning paradigm that explicitly exploits this important similar-to-different (S2D) human learning strategy, and implement it based on three algorithms (C4.5, CART and LR) efficiently. The framework not only analyzes the datasets prior to implementation, but also carefully allows classifier to have a systematic study so as to mimic the human training technique designed for efficient learning. Experimental results show that the method outperforms the state of art methods in terms of learning capability and breaks through the gap between human and machine learning. In fact, the proposed method similar-to-different (S2D) strategy may also be useful in efficient learning of realworld complex and high dimensional datasets, especially which are very typical to learn with traditional classifiers.

Keywords- Data Mining, Classification; learning strategy; Similar-to-Different (S2D).

10. Paper 31051347: Implementation of Back Propagation Algorithm For Estimation of Stress and Strain of Alloy Wheel (pp. 68-72)

R.I. Rajidap Neshtar, Research Scholar, Department of Mechanical Engineering, Vinayaka Missions University, Salem, India.
Dr. S. Purushothaman, Professor, Department of Mechanical Engineering, PET Engineering College, Tirunelveli District-627117, India

Abstract — This paper presents estimation of stress and strain of a Rapid prototype product using artificial neural network (ANN). Back propagation algorithm is used to train the ANN topology. 3D model of alloy wheel is developed by using PROE. The model is analyzed using ANSYS to find the Von Mises stress and equivalent strain. The algorithm is trained using 15 values in the input layer of the ANN topology and two values in the output layer: stress and strain that are to be estimated during the testing stage of BPA algorithm. The number of nodes in the hidden layer for BPA varies depending upon the weight updating equations.

Keywords - Back Propagation Algorithm, Finite Element Method, Structural Analysis, Alloy Wheel, Mean Squared Error.

11. Paper 31051348: An Approach to Reveal Website Defacement (pp. 73-79)

Rajiv Kumar Gurjwar, Computer Science Engineering Department, MANIT, Bhopal (M.P.), INDIA
Divya Rishi Sahu, Computer Science Engineering Department, MANIT, Bhopal (M.P.), INDIA
Deepak Singh Tomar, Computer Science Engineering Department, MANIT, Bhopal (M.P.), INDIA

Abstract — Due to ad-hoc nature of web application development and design complexity of web application, it is difficult to attain fool proof web security. In recent years invaders defaced several web sites by projecting techniques such as phishing, code injection etc. In the web defacement attack the invader changes the visual

appearance of the webpage. The business competitor, Insurgent and extremist groups defame the reputation of the organizations and mislead public through these types of attacks. Manual monitoring and scrutinizing these attacks on web sites is a time consuming and tedious task for law enforcement agencies. Hence there is a need to develop a system which effectively monitors the content of web sites and automatically generate alarm for any suspicious or threatening activity. In this work a prototype system is developed to scrutinize and detects the defacement activities automatically. At first phase web contents are preprocessed and stored in the web domain dictionary. Second phase checked integrity of web contents through CRC32, MD5, SHA 512, PSNR and SSIM techniques. The developed system successfully scrutinizes the web defacement attacks and it would be helpful for web administrator to monitor the web defacement attacks.

Keywords - web security; website defacement; internet security.

12. Paper 31051309: Java Card For PayTV Application (pp. 80-85)

Pallab Dutta, Team Leader, Centre For Development Of Telematics, Electronic City, Bangalore, India.

Abstract — Smart cards are widely used along with PayTV receivers to store secret user keys and to perform security functions to prevent any unauthorized viewing of PayTV channels. Java Card technology enables programs written in the Java programming language to run on smart cards. Smart cards represent one of the smallest computing platforms in use today. The memory configuration of a smart card are of the order of 4K of RAM, 72K of EEPROM, and 24K of ROM. Using Java card provides advantages to the industry in terms of ease of coding, faster time to market and faster upgrades as compared to plain smart cards . Also different applications like payTV, e-commerce, health-card can easily be implemented in a single java card as multiple applets corresponding to each application can coexists in a single java card. But there are security concerns in java cards and also the performance issues. In this paper, we analyse the suitability of using Java card for PayTV applications as part of conditional access system in place of plain smart cards.

Keywords- Smart Card, Java Card, PayTV, Conditional Access System (CAS), Cryptography

13. Paper 31051320: Implementation of Omni Directional Cognitive Memory in Office Mobile Robot (pp. 86-92)

S. Sivagnana sundari, Research Scholar, Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, INDIA.
Dr. C. Vijayalakshmi, Professor, School of Advance Sciences, Department of Mathematics Division,VIT University, Chennai, INDIA.

Abstract - This paper presents a method for building omni directional memory from two dimensional memory storage. The omni directional memory is implemented in the office mobile robot. Time series method is used to estimate the next position of the robot based on the stored memory. Images and sounds are collected in the office environment to store expert database in the memory of the robot. A section of the image frames taken in the corridor and how the image is associated in the omni directional memory is shown. Based on the information in the memory and the speed with which the robot is moving, the method of predicting the next position by time series method is discussed.

Keywords: Omni directional, Mobile robot, Two dimensional memory, Time series method

14. Paper 31051329: Suggest an Aspect-Oriented Design Approach for UML Communication Diagram (pp. 93-98)

Mohammed F. Nather, Software Engineering Dept., University of Mosul, Mosul, Iraq
Dr. Nada N. Saleem, Software Engineering Dept., University of Mosul, Mosul, Iraq

Abstract - More and more works are done on the design of the Unified Modeling Language (UML) which is designed to help us for modeling effective object oriented software , Existing Object-Oriented design methods are not mature enough to capture non-functional requirement such as concurrency, fault tolerance , distribution and persistence of a software approach. Our approach proposed to use aspect-oriented software development (AOSD) mechanisms to solve the issues for interactions of the communication diagram in UML that support only the Object-Oriented mechanisms ,thus AOSD allow to design programs that are out of reach of strict Object-Orientation and could possibly improve the structures and implementations.

Keywords-Aspect-Oriented Programming (AOP), Unified Modeling Language (UML) Aspect Oriented Software Development (AOSD), Software Engineering (SE), Separation of Concerns (SoC)

15. Paper 31051337: Satellite Image Classification Methods and Landsat 5TM Bands (pp. 99-103)

Jamshid Tamouk, Nasser Lotfi, Mina Farmanbar
Department of Computer Engineering, EMU University, Famagusta, North Cyprus

Abstract — This paper attempts to find the most accurate classification method among parallelepiped, minimum distance and chain methods. Moreover, this study also challenges to find the suitable combination of bands, which can lead to better results in case combinations of bands occur. After comparing these three methods, the chain method over perform the other methods with 79% overall accuracy. Hence, it is more accurate than minimum distance with 67% and parallelepiped with 65%. On the other hand, based on bands features, and also by combining several researchers' findings, a table was created which includes the main objects on the land and the suitable combination of the bands for accurately detecting of land cover objects. During this process, it was observed that band 4 (out of 7 bands of Landsat 5TM) is the band, which can be used for increasing the accuracy of the combined bands in detecting objects on the land.

Keywords: parallelepiped, minimum distance, chain method, classification, Landsat 5TM, satellite band

16. Paper 31051343: A Survey of Satellite Imagery Classification with Different Approaches (pp. 104-108)

Dr. Ghayda A. Al-Talib, Dept. of Computer Sciences, College of Mathematics and Computer sciences, University of Mosul, Mosul, Iraq
Ekhlas Z. Ahmed, Dept. of Computer Sciences, College of Mathematics and Computer sciences, University of Mosul, Mosul, Iraq

Abstract — This paper, proposes a new classification method that uses Hidden Markov Models (HMM s) to classify remote sensing imagery by exploiting the spatial and spectral information. When applying unsupervised classification to remote sensing images it can provide more useful and understandable information. Experiments shows that other clustering scheme like traditional k-means does not performs well because it does not take into account the spatial dependencies. Experiments are conducted on a set of multispectral satellite images. Proposed algorithm is verified for simulated images and applied for a selected satellite image processing in the MATLAB environment.

Index Terms — Hidden Markov Models(HMM), land cover, multispectral satellite images, unsupervised classification.

17. Paper 23021317: Secure Routing in UAV (pp. 109-122)

Ahmed Refaat Sobhy, Rowayda.A.Sadek, Atalla Hashad
Arab Academy for Science & Technology & Maritime Transport, College of Engineering & Technology, Cairo, Egypt

Abstract — The field of UAV has gained an important part of the interest of researchers and become very popular in last few years. Focusing in the routing protocols used in UAV's systems in order to obtain a secure routing protocol this paper presents the effect of DOS attack on two different types of routing protocols , proactive and reactive routing protocols. The proactive routing protocol is represented by OLSR routing protocol and the reactive routing protocol is represented by AODV , TORA routing protocols . in addition the performance metrics of ordinary routing protocols (OLSR , AODV , TORA) are compared in three different scenarios implemented by using Opnet simulator. The simulation results will show the performance impact of security implements into reactive & proactive protocols after implementations of Authentication & encryption algorithms. The main concern of this paper is to propose an efficient and secure routing protocol to the UAV.

18. Paper 31031339: Survey on Internet-based Mobile Ad Hoc Networking (pp. 123-128)

Omung Goyal, Swati Jaiswal, Prateek Poste
Dept. of CSE, UIT RGPV, Bhopal

Abstract - Internet-based Mobile Ad Hoc Networking (MANET) is an emerging technology that supports self-organizing mobile networking infrastructures. This is expected to be of great use in commercial applications for the next generation internet users. A number of technical challenges are faced today due to the heterogeneous, dynamic nature of this hybrid MANET. A new hybrid routing scheme AODV ALMA is proposed, which act simultaneously combining mobile agents to find path to the gateway and on-demand distance vector approach to find path in local MANET is one of the unique solution. An adaptive gateway discovery mechanism based on mobile agents making use of pheromone value, pheromone decay time and balance index is used to estimate the path and next hop to the gateway. The mobile nodes automatically configure the address using mobile agents first selecting the gateway and then using the gateway prefix address. The mobile agents are also used to track changes in topology enabling high network connectivity with reduced delay in packet transmission to Internet.

Keywords: Hybrid MANET, mobile agents, AODV ALMA, adaptive gateway discovery.

3-SAT Problem: A New Memetic-PSO Algorithm

Nasser Lotfi
Department of Computer Engineering
EMU University
Famagusta, North Cyprus

Jamshid Tamouk
Department of Computer Engineering
EMU University
Famagusta, North Cyprus

Mina Farmanbar
Department of Computer Engineering
EMU University
Famagusta, North Cyprus

***Abstract*—3-SAT problem is of great importance to many technical and scientific applications. This paper presents a new hybrid evolutionary algorithm for solving this satisfiability problem. 3-SAT problem has the huge search space and hence it is known as a NP-hard problem. So, deterministic approaches are not applicable in this context. Thereof, application of evolutionary processing approaches and especially PSO will be very effective for solving these kinds of problems. In this paper, we introduce a new evolutionary optimization technique based on PSO, Memetic algorithm and local search approaches. When some heuristics are mixed, their advantages are collected as well and we can reach to the better outcomes. Finally, we test our proposed algorithm over some benchmarks used by some another available algorithms. Obtained results show that our new method leads to the suitable results by the appropriate time. Thereby, it achieves a better result in compared with the existent approaches such as pure genetic algorithm and some verified types.**

Keywords: 3-SAT problem; Particle swarm optimization; Memetic algorithm; Local search.

I. INTRODUCTION

3-SAT problem is of great importance to achieve higher performance in many applications. This paper presents a new hybrid evolutionary algorithm for solving this satisfiability problem. 3-SAT problem has the huge search space and it is a NP-hard problem [1]. Therefore, deterministic approaches are not recommended for optimizing of these functions with a large number of variables [2]. In contrast, an evolutionary approach such as PSO may be applied to solve these kinds of problems, effectively. There exist a few genetic algorithms for solving 3-SAT problem. The representation of a problem solution, encoding scheme, highly affects the speed of genetic algorithms. The primary difference amongst genetic algorithms is the chromosomal representation, Crossover scheme, mutation Scheme and Selection strategy. Evolutionary optimization algorithms mainly encode the value of variables as string of bits. But the reported results show that they alone cannot approach to optimal point sufficiently. Also these algorithms spend more time to get these results. The performance of an evolutionary algorithm is often sensitive to the quality of its initial population [2]. A suitable choice of the initial population may accelerate the convergence rate of evolutionary algorithms because, having an initial population with better fitness values, the number of generations required to get the final individuals, may reduce. Further, high diversity in the population inhibits early convergence to a locally optimal solution [2]. In our produced way we observe this rule and produce the initial particles intelligently. The initial population of particles is usually generated randomly. The "goodness" of the initial population depends both on the average fitness (that is, the objective function value) of individuals in the population and the diversity in the population [2]. Losing on either count tends to produce a poor evolutionary algorithm. As it is described in the future Sections, by creating an initial particles as intelligently, the convergence rate of our proposed algorithm is highly accelerated.

Previous genetic algorithms used the simple operators to produce new population that have weak diversity [2]. In our proposed algorithm we have used a suitable way to represent particles that have several advantages. Important one is that the count of population to reach the final population reduced, because the algorithm starts by the convenient initial particles. Finally, it achieves a better value in comparison with the existing approaches such as genetic algorithm.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The remaining parts of this paper are organized as follows: In Section 2, the 3-SAT problem is outlined. Section 3 presents a structure of PSO algorithms. In Section 4, the proposed algorithm based on PSO and Memetic algorithms are described. A practical evaluation of the proposed optimization algorithm is presented in Section 5. Finally, section 6 states the conclusion and future works.

II. 3-SAT PROBLEM

In this section, description of the multivariable function is presented. The SAT problem is one of the most important optimization combinatorial problems because it is the first and one of the simplest of the many problems that have been proved to be NP-Complete [3]. A Boolean satisfiability problem (SAT) involves a Boolean formula F consisting of a set of Boolean variables $x_1, x_2, ..., x_n$. The formula F is in conjunctive normal form and it is a conjunction of m clauses $c_1, c_2, ..., c_m$. Each clause c, is a disjunction of one or more literals, where a literal is a variable x_j or its negation. A formula F is satisfiable if there is a truth assignment to its variables satisfying every clause of the formula, otherwise the formula is unsatistiable. The goal is to determine a variable x assignment satisfying all clauses [4].

For example, in the formula below p1, p2, p3 and p4 are propositional variables. This formula is named CNF.

$$(p_1 \vee p_2 \vee \neg p_3) \wedge (\neg p_1 \vee p_2 \vee p_3) \wedge (\neg p_1 \vee \neg p_2 \vee p_3) \wedge (p_1 \vee \neg p_3 \vee p_4)$$

The class k-SAT contains all SAT instances where each clause contains exactly k distinct literals. While 2-SAT is solvable in polynomial time, k-SAT is NP-complete for $k \geq 3$ [5]. The SATs have many practical applications (e.g. in planning, in circuit design. in spin-glass model. in molecular biology ([6], [7], [8]) and especially many applications and research on the 3-SAT is reported. Many exact and heuristic algorithms have been introduced.
As described above in Section 1, 3-SAT optimization problem is a NP-hard problem which can be best solved by applying an evolutionary optimization approaches. In the following, we consider the PSO and Memetic algorithms and using them to solve this problem.

III. PARTICLE SWARM OPTIMIZATION AND MEMETIC ALGORITHMS

Particle swarm optimization (PSO) [9] is a population based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by the social behavior of birds. The algorithm is very simple but powerful. A "swarm" is an apparently disorganized collection (population) of moving individuals that tend to cluster together while each individual seems to be moving in a random direction. We also use "swarm" to describe a certain family of social processes. The PSO approach utilizes a cooperative swarm of particles, where each particle represents a candidate solution, to explore the space of possible solutions to an optimization problem. Each particle is randomly or heuristically initialized and then allowed to 'fly' [9]. At each step of the optimization, each particle is allowed to evaluate its own fitness and the fitness of its neighboring particles. Each particle can keep track of its own solution, which resulted in the best fitness, as well as see the candidate solution for the best performing particle in its neighborhood. At each optimization step, indexed by t, each particle, indexed by i, adjusts its candidate solution (flies) according to (1) and Figure 1 [10].

$$\vec{v}_i(t+1) = \vec{v}_i(t) + \phi_1(\vec{x}_{i,p} - \vec{x}_i) + \phi_2(\vec{x}_{i,n} - \vec{x}_i)$$
$$\vec{x}_i(t+1) = \vec{x}_i(t) + \vec{v}_i(t+1) \quad (1)$$

Figure1. Compute the particles's new location

First equation in (1) may be interpreted as the 'kinematic' equation of motion for one of the particles (test solution) of the swarm. The variables in the dynamical system of first equation are summarized in Table1 [10].

TABLE I. VARIABLES USED TO EVALUATE THE DYNAMICAL SWARM RESPONSE

Parameter	Description
$\vec{v}_i$	The particle velocity
$\vec{x}_i$	The particle position (Test Solution)
t	Time
Φ_1	A uniform random variable usually distributed over [0,2]
Φ_2	A uniform random variable usually distributed over [0,2]
$\vec{x}_{i,p}$	The particle's position (previous) that resulted in the best fitness so far
$\vec{x}_{i,n}$	The neighborhood position that resulted in the best fitness so far

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 2 shows the Algorithm pseudo code of PSO Generally.

```
I ) For each particle:
                Initialize particles.
II ) Do:
    a) For each particle:
        1) Calculate fitness value
        2) If the fitness value is better than the best Fitness
           value  (pBest) in history
        3) Set current value as the new pBest
      End
    b) For each particle:
        1) Find in the particle neighborhood, the particle With
           the best fitness
        2) Calculate particle velocity according to the
           Velocity equation
        3) Apply the velocity constriction
        4) Update particle position according to the
           Position equation
        5) Apply the position constriction
      End
  While maximum iterations or minimum error criteria is not attained.
```

Figure 2. The PSO Algorithm pseudo code.

The combination of Evolutionary Algorithms with Local Search Operators that work within the EA loop has been termed “Memetic Algorithms”. Term also applies to EAs that use instance specific knowledge in operators. Local search is the searching of best solution among adjacent solutions that replace population members with better than. Pivot rule in the memetic algorithms have two types. At first type the search stopped as soon as a fitter neighbor is found (Greedy Ascent) and at second type the whole set of neighbors examined and the best neighbor found (Steepest Ascent). Figure 3 shows the pseudo code for local search [11].

```
Begin
 /* given a starting solution i and a  neighborhood function*/
 Set best =i ;
 Set iteration =0;
 Repeat until (depth condition is satisfied )  DO
   Set count =1;
   Repeat until (pivot rule is satisfied) DO
     Generate the next neighbor j ϵ n(i)
     Set count =count+1;
     IF (f(j) is better than f (best) THEN
        Set best =j;
     FI
   OD
   Set i=best
   Set iteration =iteration+1
 OD
```

Figure 3. The local search pseudo code

It has been shown that the memetic algorithms are faster and more accurate than GAs on some problems, and are the “state of the art” on many problems. Another common approach would be to initialize population with solutions already known, or found by another technique (beware, performance may appear to drop at first if local optima on different landscapes do not coincide) [11].

IV. A NEW MEMETIC PSO TO SOLVE 3-SAT PROBLEM

To understand the algorithm, it is best to imagine a swarm of birds that are searching for food in a defined area - there is only one piece of food in this area. Initially, the birds don't know where the food is, but they know at each time how far the food is. Which strategy will the birds follow? Well, each bird will follow the one that is nearest to the food [8].

PSO adapts this behavior and searches for the best solution-vector in the search space. A single solution is called particle. Each particle has a fitness/cost value that is evaluated by the function to be minimized, and each particle has a velocity that directs the "flying" of the particles. The particles fly through the search space by following the optimum particles [8].

The algorithm is initialized with particles at random positions, and then it explores the search space to find better solutions. But in our proposed memetic-PSO algorithm, the initial population is not produce quite random. We must produce initial population with better quality than random type. In our proposed algorithm we combine PSO, Memetic and Local search algorithms to collect their advantages in a new algorithm. To attain this population we produce 1000 particle and then select the 100 better particles among them. Or in other words, we produce initial particles by heuristic to have better swarm. Each particle represented by the binary array inclusive just 0 and 1. Length of this array is equal to number of propositional variables. For a CNF with 32 variables, we can assume the length equal to 32. An example of the particle is given in Figure 4. In this particle, the values of first and last variables are TRUE and FALSE respectively.

```
1110011011001100  0110101110000010
```

Figure 4. A chromosome created by memetic approach

In the every iteration, each particle adjusts its velocity to follow two best solutions. The first is the cognitive part, where the particle follows its own best solution found so far. This is the solution that produces the lowest cost (has the highest fitness). This value is called pBest (particle best). The other best value is the current best solution of the swarm,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

i.e., the best solution by any particle in the swarm. This value is called gBest (global best). In the 3-SAT problem, we can not use the introduced PSO formulas, because the solutions or particles in this problem are binary. Hence we must use another form of PSO named by Binary PSO. In the binary PSO the formulas we can use are as following. Then, each particle adjusts its velocity and position with the equations below in Figure 5.

$$g(v_{id}) = \begin{cases} v_{\max}, & v_{id} > v_{\max} \\ v_{id}, & -v_{\max} \le v_{id} \le v_{\max} \\ -v_{\max}, & v_{id} < -v_{\max} \end{cases}$$

$$sig(v_{id}) = \frac{1}{1+\exp(-v_{id})}$$

$$v_{id} = g\left(\omega v_{id} + c_1 R_{id}\left(p_{id} - x_{id}\right) + c_2 r_{id}\left(p_{gd} - x_{id}\right)\right)$$

$$x_{id} = \begin{cases} 1, rand < sig(v_{id}) \\ 0, otehrwise \end{cases}$$

Figure 5. Velocity and position adjustent in binary PSO

In these formulas, v_{id} and x_{id} are the new velocity and position respectively, P_{id} and P_{gd} are Pbest and Gbest, R_{id} and r_{id} are even distributed random numbers in the interval [0, 1], and c_1 and c_2 are acceleration coefficients. The c_1 is the factor that influences the cognitive behavior, i.e., how much the particle will follow its own best solution and c_2 is the factor for social behavior, i.e., how much the particle will follow the swarm's best solution.

The algorithm can be written as follows in Figure 6 [8]:

1. Initialize each particle with a random velocity and random position.
2. Calculate the cost for each particle. If the current cost is lower than the best value so far, remember this position (pBest).
3. Choose the particle with the lowest cost of all particles. The position of this particle is gBest.
4. Calculate, for each particle, the new velocity and position according to the above equations.
5. Repeat steps 2-4 until maximum iteration or minimum error criteria is not attained.

Figure 6. Binary PSO Algorithm

This is a quite simple algorithm, but not sufficiently. In our new approach in order to produce high quality particles and having sufficient power, we add memetic approach again. After producing a population we use local search to each particle and improve that's quality. In other words, we use local search algorithm in the each iteration to replace particles by better neighbors. So each particle could improve itself and helps to speedy convergence to optimal point.

The quality of each particle is simply computed. Fitness value or quality of a particle is equal to the number of elements in CNF which the particle makes them TRUE or FALSE. Being TRUE or FALSE depends on our objective.

V. EXPERIMENTAL RESULTS

In this section, the performance results and comparison of our proposed algorithm is presented. Our proposed algorithm is compared with the results of some existent algorithm [12, 13]. The comparison is made by applying our algorithm to the some famous CNFs presented in related papers. It is observed that the proposed algorithm results in better than other algorithms and it produces the better outcomes. However we don't compare our algorithm to another deterministic algorithm, because 3-SAT problem is NP-hard and Deterministic approaches are not applicable in this context. At first, we present the results of our proposed memetic PSO algorithm on random produced CNFs. Table below shows the obtained results.

TABLE II. RESULTS OVER RANDOM PRODUCED CNF'S

Variable Number	Closure Number	Result	Validity	Generations
36	12	CNF is satisfiable	Valid	100
33	65	7 Closure is not satisfiable	Valid	200
62	74	CNF is satisfiable	Valid	120
100	100	CNF is satisfiable	Valid	150
80	50	CNF is satisfiable	Valid	80
50	50	1 Closure is not satisfiable	Valid	200
93	77	CNF is satisfiable	Valid	134
83	32	CNF is Satisfiable	Valid	236
35	59	CNF is Satisfiable	Valid	176
43	90	9 Closure is not satisfiable	Valid	200
26	79	3 Closure is not satisfiable	Valid	200
88	57	CNF is Satisfiable	Valid	109
91	92	CNF is Satisfiable	Valid	167
98	56	1 Closure is not satisfiable	Valid	200
45	78	CNF is Satisfiable	Valid	111
78	100	CNF is Satisfiable	Valid	136

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Here we consider the sample CNF generated randomly with 100 variables and 100 Closures. Figure 7 shows the first population generated by memetic algorithm that's including the better particles. Variation between particles can be seen.

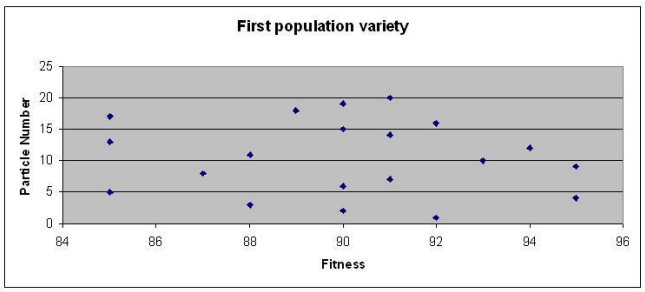


Figure 7. First population generated by memetic algorithm

The evolution of the chromosomes, while applying our proposed evolutionary algorithm on the mentioned example, is shown below in Figure 8. We can see that the fitness of best particle is gradually improved generation by generation.

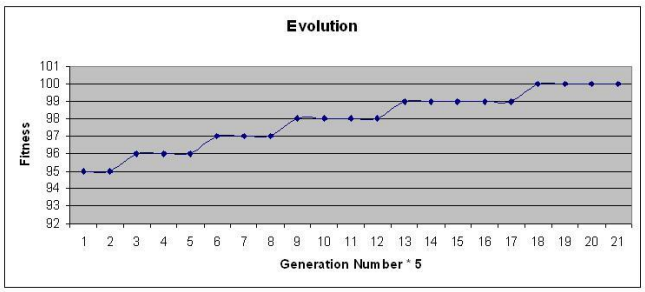


Figure 8. Evolution of particles

Also, in order to demonstrate the stability of the results obtained in the above example, the results obtained by twenty runs of the algorithm are compared in Figure 9. We can see that all 100 closures are satisfied in all 20 runs.

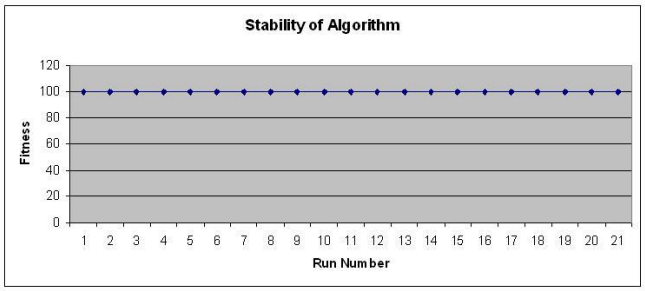


Figure 9. Best fitness obtained in 100 generations and 20 runs

We continue our evaluating using two existent well known algorithms to solve this problem [12, 13].

At first, we evaluate the performance of our proposed algorithm on several classes of satisfiable and unsatisfiable benchmark instances and compare it with GASAT [12] and with WALKSAT [14], one of the well-known incomplete algorithms for SAT, and with UNIWALK [15], the best up-to-now incomplete randomized solver presented to the SAT competitions [12]. Two classes of instances are used: structured and random instances. Structured instances are aim-100-1_6-yes1-4 (100 variables and 160 clauses), aim-100-2_0-yes1-3 (100 variables and 200 clauses), math25.shuffled (588 variables and 1968 clauses), math26.shuffled (744 variables and 2464 clauses), color-15-4 (900 variables and 45675 clauses), color-22-5 (2420 variables and 272129 clauses), g125.18 (2250 variables and 70163 clauses) and g250.29 (7250 variables and 454622 clauses).

Also, the random instances are glassy-v399-s1069116088 (399 variables and 1862 clauses), glassy-v450-s325799114 (450 variables and 2100 clauses), f1000 (1000 variables and 4250 clauses) and f2000 (2000 variables and 8500 clauses) [12]. Two criterions are used to evaluation and comparison. First one is the success rate (%) which is the number of successful runs divided by the total number of runs. The second criterion is the average running time in second. We have tried to use same computer and hardware for running [12]. Tables below show the comparison between these four algorithms. If no assignment is found then the best number of false clauses is written between parentheses.

TABLE III. STRUCTURED INSTANCES

Benchmarks	Our Algorithm	GASAT	WALKSAT	UNITWALK
aim-100-1_6-yes1-4	100% 27.19	10% 84.53	(1 clause)	100% 0.006
aim-100-2_0-yes1-3	100% 14.32	100% 20.86	(1 clause)	100% 0.0019
math25.shuffled	(3 clauses)	(3 clauses)	(3 clauses)	(8 clauses)
math26.shuffled	(2 clauses)	(2 clauses)	(2 clauses)	(8 clauses)
color-15-4	100% 358.43	100% 479.248	(7 clauses)	(16 clauses)
color-22-5	(5 clauses)	(5 clauses)	(41 clauses)	(51 clauses)
g125.18	100% 281.455	100% 378.660	(2 clauses)	(19 clauses)
g250.29	(45 clauses)	(57 clauses)	(34 clauses)	(57 clauses)

TABLE IV. RANDOM INSTANCES

Benchmarks	Our Algorithm	GASAT	WALKSAT	UNITWALK
glassy-v399-s1069116088	(5 clauses)	(5 clauses)	(5 clauses)	(17 clauses)
glassy-v450-s325799114	(10 clauses)	(8 clauses)	(9 clauses)	(22 clauses)
F1000	100% 34.45	100% 227.649	100% 9.634	100% 1.091
F2000	100% 19.94	(6 clauses)	100% 21.853	100% 17.169

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

As we can see in tables, our proposed algorithm works better than others in overall and is more efficient from the performance view.

VI. CONCLUSIONS AND FUTURE WORKS

3-SAT problem is NP-hard and can be considered as an optimization problem. To solve this NP-hard problem, non-deterministic approaches such as evolutionary algorithms are quite effective.

Values of propositions can be best encoded as a binary array. The objective of evolutionary algorithms can be to maximize the number of valid DNF elements in CNF. In this way, the fitness of each particle in a population depends on the value of DNF elements. We used PSO approach based on memetic algorithms to solve this problem that is better than existent approaches.

The other kind of this problem is multi objective SAT problem that's more important. Multi-objective optimization problems consist of several objectives that are necessary to be handled simultaneously. Such problems arise in many applications, where two or more, sometimes competing and/or incommensurable, objective functions have to be minimized concurrently. It's possible to use evolutionary approaches to solve such problems [10].
Multivariable SAT problem can be defined in the form of multi-objective optimization problem. In this form, we deal with m formulas, each representing a different objective. The goal is to satisfy the maximum number of clauses in each formula. For solving this problem, we can extend our proposed memtic PSO to the multi-objective problems solver form. Hence, the set of non-dominated solutions must be found for this kind of problem.

REFERENCES

[1] Garey M. R. and Johnson D. S., "Computers and Intractability: A Guide to the Theory of NP-Completeness", W.H. Freeman and Company, 1979.
[2] S.Parsa, S.Lotfi and N.Lotfi, "An evolutionary approach to task graph scheduling, ICANNGA 2007", Springer-verlogBerlin Heidelberg 2007, LNCS 4431, pp 110-119,2007.
[3] J.S, D.Garey and R.Michael, "Computers and Interactibility", Computer Science / Mathematics, W.H. freeman and company, March, 1991.
[4] I. Borgulya, "An evolutionary framework for 3-SAT problems", 25th Int. Conf. Information Technology Interfaces IT1 2003, June 16-19, 2003.
[5] Carey M., Johnson D., Computers and Intractability: a Guide to the Theory of NP-completeness, Freeman. San Francisco. CA, 1997.
[6] Du D.,Cu J., Pardalos P., (eds). Satisfiability Problem: Thcoiy and Applications. (1 997) Vol. 35. DlMACS Series in Discrete Mathematics and Theoretical Computer Science, AMs, Providence, Rhode Island.
[7] Crisanti A., Lcuzzi L., Parisi G., Tlic 3-SAT problem with largc nuinbcr of clauses in the co-replica symmetry breaking scheme. Phys. A: Math. CC11 35 (2002) 481-497.
[8] Hagiya M., Rose J.A., Komiya K., Sakamoto K.. Complcxity analysis of the SAT enginc: DNA algorithms as probabilistic algorithms. Theoretical Computer Science 287 (2002) 59-71.
[9] Eberhart, R.C. and J. Kennedy. A new optimizer using particle swarm theory. in Proceedings of the Sixth International Symposium on Micro Machine and Human Science. 1995. Nagoya, Japan.
[10] J.Tillet, T.M.Rao, F.Sahin and R.Rao, "Darwinian particle swarm optimization", University of Rochester, newyork, USA, 2008.
[11] A.E.Eiben, "Introduction to Evolutionary Computing", Springer, 2007.
[12] J.K.HAo, F.Lardeux and F.Saubion, "Evolutionary computing for the satisfiability problem", international conference on Applications of evolutionary computing, Springer-verilog, Berlin, 2003.
[13] J.M.Howe and A.King, "A Pearl on SAT solving in prolog", *Tenth Int*ernational Symposium on Functional and Logic Programming, Lecture Notes in Computer Science, page 10. Springer-Verlag, April 2010.
[14] B.Selman, H.A.Kautz and B.Kohen, "Noise strategies for improving local search", In proc of AAAI, Vol.1, pages 337-343, 1994.
[15] E.A.Hirsch and A.kojevnikov, "A new SAT solver hat uses local search guided by unit clause elimination", PDMI preprint 2001, Petersburg, 2001.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Design a Data Model Diagram from Textual Requirements

Thakir N. Abdullah
Msc. Student, Software Engineering Dept.
University of Mosul
Mosul, Iraq

Dr. Nada N. Saleem
Asst. Prof., Software Engineering Dept.
University of Mosul
Mosul, Iraq

***Abstract*— This paper try to automate the process of designing data model diagram (Entity Relationship Diagram) from textual requirements. It focuses on the very early stage of the database development which is the stage of user requirement analysis. It is supposed to be used between the requirements determination stage and analysis. The approach provides the opportunity of using natural language text documents as a source and extract knowledge from textual requirements for generation of a conceptual data model. The system performs information extraction by parsing the syntax of the sentences and semantically analyzing their content.**

***Index Terms*— natural language processing, textual requirements, conceptual data modeling, heuristic rules.**

I. INTRODUCTION

Discovering the knowledge required to design a data model from user requirements is an elaborate process because: (1) Users may say words that we don't need them to design a data model diagram. (2) Users may give conflicted or incomplete requirements. these (unneeded words or incomplete requirements) aren't discovered until designing a data model diagram, So that we need to contact the user again.

This paper creates a tool to extract knowledge from textual requirements for creating Entity Relationship Diagram (ERD) based on Natural language processing techniques to break up or end that elaborate process and helps to detect defects and provides traceability between sentences and the ERD elements.

ERD was used for a long time as communication tool with the users and as a blueprint for the data modelers. As database design get more complicated, the diagram becomes no more ideal communication tool with the users, because users don't have technical information and the modeler need details. It was found that between 93% and 95% of all the user requirements in industrial practice were written in natural language[1]. So that the textual requirements is ideal for communication (if it is free from conflicts and ambiguities) and ERD is ideal for data modeler. In this paper the NLP techniques is used to produce data model diagram from textual requirements so that the textual requirements will stay as a communication tool and the data model produced from it will be the blueprint for modelers. This tool will automate the process of transfer textual requirements to data model diagram.

Because of the success of software engineering in making many steps of constructing software semi automatic (from design to generate code to test it) as seen in [2]. so that this paper will try to increase that success to include software process automation by textual requirement analysis step. The automated development of software has the potential to reduce human error in the creation of code that must meet precise syntax and other constraints. It has the potential to produce similar or better software than that produced 'by hand' by relatively scarce skilled software development talent, potentially reducing costs. Automated development may lead to a greater use of standardized components, thus increasing software reliability and decreasing the future maintenance costs of software. Finally, automation may reduce the number of less-interesting, more-mechanical tasks software developers are required to perform, thus freeing them to focus on tasks that require more creativity[2].

the modeling process is so time consuming [3] (see Figure 1).

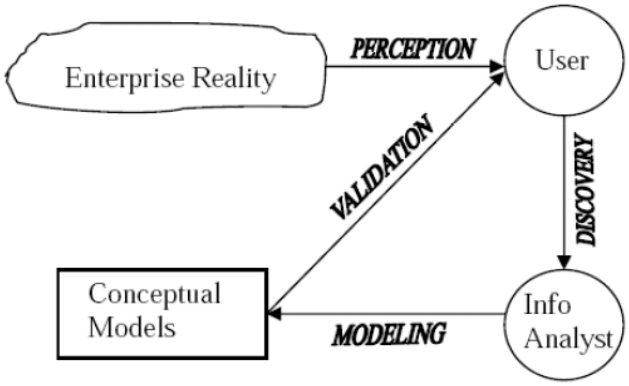


Figure 1.
A process model of information requirements.

The loop discovery-modelling-validation is time consuming and the textual requirements may contain conflicted requirements that will not discovered until modelling step, and trying to break up or end that loop is the goal of this paper.

Information systems development suffers from two widely acknowledged problems:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

1) an applications backlog: whereby demand for applications exceeds resources available for its satisfaction.
2) a requirements analysis problem: This is often manifested as a maintenance problem, whereby resources that could be put into reducing the applications backlog are instead devoted to correcting faults in delivered systems. Most such faults are traceable to erroneous specifications, resulting from a failure to establish user requirements correctly [4], and that is the reason for automated software process in natural languages.

II. Related Works and our Contribution

A. Related Works

Recently many papers try to make the process of conceptual modeling automatic as in [5]. Aforementioned paper try to get UML object oriented design for textual requirements sentences by proposing heuristic rules to map part of speech tags to UML elements. Chen proposed eleven high level heuristic rules to map basic constructs of English sentences into ERDs [6]. Chen mentioned that they are better viewed as "guidelines" since it is possible to find counter examples to them. Reference [7] creates detailed rules as Chen's rules for mapping and it is more specific. Another proposed a high level heuristic rules for extracting Entity relation tuples based on high level Stanford typed Dependencies and link parser [8]. Heuristics-based approaches are the best-known approaches to NLP-based conceptual data modeling because heuristics, often guided by common sense, provides good but not necessarily optimal solutions to many difficult problems such as automated conceptual data modeling where precise algorithmic solutions are not available[1].

Reference [9] proposed six domain independent modeling rules. The author thought that there is always trade off in design so that not all previous proposed rules can work together because some rules are conflicting. Reference [10] proposed an intermediate level for requirements representation, an interlingua connecting the natural language level of the end user and conceptual model level produced by engineers.

B. Our Contribution

we proposed heuristic rules based on Penn Treebank in order to extract information from textual requirements to construct ERD. Since there are many heuristic rules proposed and some of these heuristic rules depend either on basic constructs of English sentences and these usually used as a guidelines for human, or on specific parser's output such as Link type or Stanford typedDependencies. our proposed heuristic rule depend on Penn Treebank since Penn Treebank produced by many parsers such as Link parser, Stanford parser, openNLP parser and others.

III. Rsearch Approach

we proposed heuristic rules based on Penn Treebank in order to extract information from textual requirements to construct ERD. First, we extract a largest noun phrase that may contain other noun phrases in it and we called it most upper noun phrase, then based on the number of most upper noun phrases in the sentence, our proposed heuristic rules applied. before giving rules we will explain how to extract most upper noun phrase and how to extract verb of the relation:

A. Extracting Most Upper Noun Phrase

The most upper noun phrase is the NP that doesn't contain the following bracket Labels of Penn Treebank (S,VP,PP) and must contain one noun at least (NN,NNP, NNS or NNPS). Example:

Each company has several plants and many employees.

```
(ROOT
 (S
  (NP (DT Each) (NN company))
  (VP (VBZ has)
   (NP
    (NP (JJ several) (NNS plants))
    (CC and)
    (NP (JJ many) (NNS employees))))
  (. .)))
```

In this sentence there are two most upper NP ,(NP (DT Each) (NN company)) and (NP (NP (JJ several) (NNS plants))(CC and)(NP (JJ many) (NNS employees))). Each most upper NP may contain more than one entity. We will extract entities and their minimum and maximum cardinality by passing these NPs to a module that implements (figure 2).

B. Extracting Verb of The Relation

To extract relationship between entities, which usually represents the verb between them, The smallest VP that contain a verb (VB, VBD, VBG, VBN, VBP or VBZ) is the relation and if a PP is one of the 'so called' direct sons of that VP then the (IN) of that PP is a part of the relation. Example:

Each person works on some projects.

```
(ROOT
 (S
  (NP (DT each) (NN person))
  (VP (VBZ works)
   (PP (IN on)
    (NP (DT some) (NNS projects))))
  (. .)))
```

The relation is 'works-on'.

C. Heuristic Rules

- Heuristic 1: If there are two most upper noun phrase in the sentence then the last VP is the relation. Example:

Each company has many plants.

```
(ROOT
 (S
  (NP (DT Each) (NN company))
  (VP (VBZ has)
   (NP (JJ many) (NNS plants)))
  (. .)))
```


http://sites.google.com/site/ijcsis/
ISSN 1947-5500

As we see (NP (DT Each) (NN company)) and (NP (JJ many) (NNS plants)) are the most upper noun phrase. The relation is (VBZ has) then a tuple <company has plants> extracted. Another Example:

Advanced courses need to be taught by professors.

```
(ROOT
  (S
    (NP (JJ advanced) (NNS courses))
    (VP (VBP need)
      (S
        (VP (TO to)
          (VP (VB be)
            (VP (VBN taught)
              (PP (IN by)
                (NP (NNS professors))))))))
    (. .)))
```

tuple <advanced-courses taught-by professors> extracted.

- Heuristic 2: If there are 3 most upper Noun phrase then
  a) If VP between 1st and 2nd NPs and another VP between 2nd and 3rd NPs then there is a relation between 1st and 2nd NPs and another relation between 1st or 2nd (the one that is its depth is lower than or equal to the depth of VP) and the 3rd NPs. Example:

The company has plants located in 40 states.

```
(ROOT
  (S
    (NP (DT the) (NN company))
    (VP (VBZ has)
      (NP
        (NP (CD 50) (NNS plants))
        (VP (VBN located)
          (PP (IN in)
            (NP (CD 40) (NNS states))))))
    (. .)))
```

Tuple <company has plants> <plants located-in states> extracted.

Each student has to take several courses and work on one project.

```
(ROOT
  (S
    (NP (DT Each) (NN student))
    (VP (VBZ has)
      (S
        (VP (TO to)
          (VP
            (VP (VB take)
              (NP (JJ several) (NNS courses)))
            (CC and)
            (VP (VB work)
              (PP (IN on)
                (NP (CD one) (NN project))))))))
    (. .)))
```

Tuples <student take courses> <student work-on project> extracted.

  b) If there is only one VP, then ternary relation between them.

Employees perform work tasks at work stations.

```
(ROOT
  (S
    (NP (NNS employees))
    (VP (VBP perform)
      (NP (NN work) (NNS tasks))
      (PP (IN at)
        (NP (NN work) (NNS stations))))
    (. .)))
```

Ternary tuple <employees perform-at work-tasks work-stations> extracted.

- Heuristic 3: If there are 4 most upper noun phrases then If VP between 1st and 2nd NP and another VP between 3rd and 4th NP then 2 relation must extract, otherwise ask the user to decompose that sentence. Example:

When a student buys books, the student may get a discount.

```
(ROOT
  (S
    (SBAR
      (WHADVP (WRB When))
      (S
        (NP (DT a) (NN student))
        (VP (VBZ buys)
          (NP (NNS books)))))
    (, ,)
    (NP (DT the) (NN student))
    (VP (MD may)
      (VP (VB get)
        (NP (DT a) (NN discount))))
    (. .)))
```

Tuples <student buys books> <student get discount> extracted.

D. Minimum and Maximum Cardinality Extraction

The previous researches maintain the problem of maximum cardinality only. Reference [1] suggested a word sequences where if they are founded then they indicate the maximum cardinality and make them one word (see table 1), in the same way for two words or more of attribute. Reference [7] proposed Heuristics to determine cardinalities:

1. Heuristic HC2: The adjective “many” or “any” may suggest a maximum cardinality. For example:
   a) “A surgeon can perform many operations.”
   b) “Each diet may be made of any number of servings.”
2. Heuristic HC3: A comparative adjective “more” followed by the preposition “than” and a cardinal number may indicate the degree of the cardinality between two entities. For example: “Each patient could have more than one operation.”

Also Reference [9] specify the maximum cardinality constraints only. In this paper we tried to specify minimum and maximum cardinality using state machine (see figure 2). For each sentence we take part of speech tags for all words of that sentence and passed it to a module that implements (figure 2) in order to extract minimum and maximum cardinality as well as the entity name, for example:

Each department can have anywhere between 1 and 10 employees and each employee has 1 and only 1 department.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

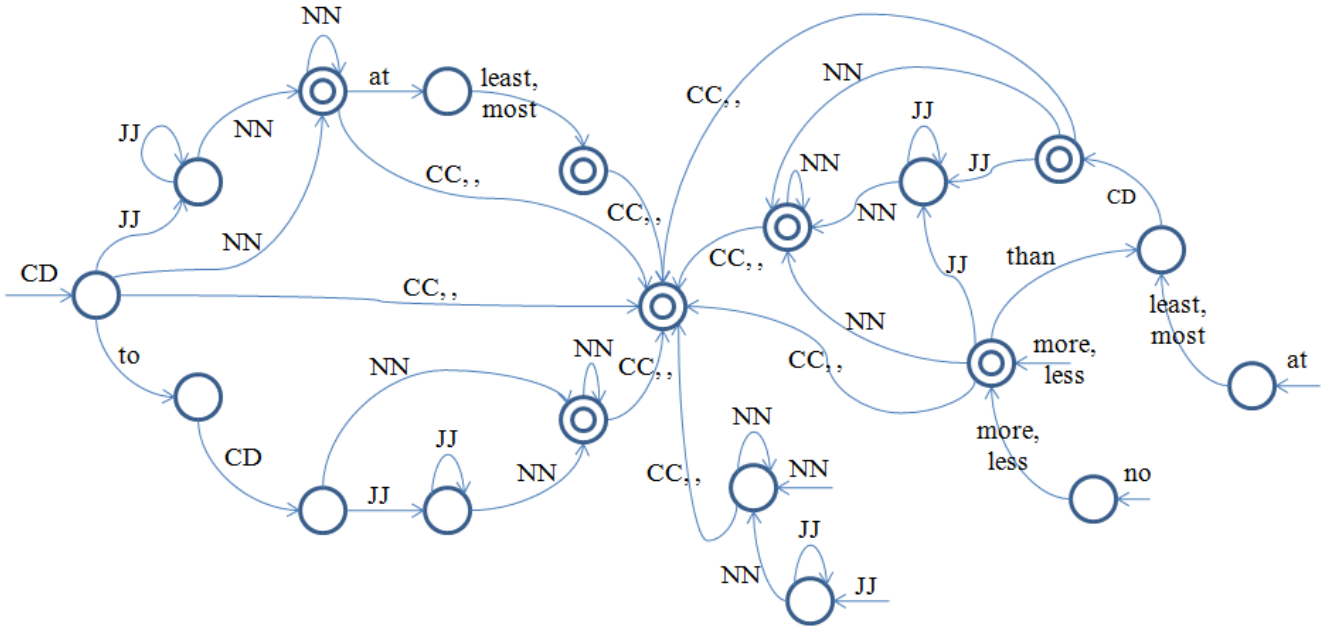


Figure 2. State machine for extraction min and max cardinality.

TABLE 1. Concatenation of domain phrases

more than one	more-than-one
at least one	at-least-one
at least two	at-least-two
one or more	one-or-more
one and only one	one-and-only-one
no more than	no-more-than
zero or more	zero-or-more
more than two	more-than-two
more than three	more-than-three
first name	first-name
zip code	zip-code
part time	part-time
full time	full-time
phone number	phone-number
email address	email-address
social security number	social-security-number
id number	id-number
year of birth	year-of-birth
mailing address	mailing-address
item number	item-number

The Part Of Speech tags of it is:

[each/DT, department/NN, can/MD, have/VB, anywhere/RB, between/IN, 1/CD, and/CC, 10/CD, employees/NNS, and/CC, each/DT, employee/NN, has/VBZ, 1/CD, and/CC, only/RB, 1/CD, department/NN, ./.]

From that POS and according to the figure 2 we extract.

Employees, min=1, max=10.

Department, min=1, max=1.

Instead of using predefined sequence of words as proposed by [1] we predicts minimum and maximum cardinality based on words around entities names.

E. Predicting Generalization

Generalization structure enables the data modelers to partition an entity type into subsets. Each subset is a part of the whole. For example, trucks, cars and van may be considered as subtypes of the supertype called vehicle. This modeling structure preserves cohesiveness in a model[11].

The data modeler usually predicts generalization for two entities or more when: (1) there are a common attributes between them and special attribute to each one or (2) there are a common relation with other entity. when a data modeler decided that there is a generalization, he tries to predict a name for supertype entity, a name that all entities under it is kind of it. That is what we try to do using WorldNet dictionary. if there are two entities or more have common attributes or relation then get coordinate terms of each noun (entity) and the shared coordinate terms between them is the supertype entity name.

IV. IMPLEMENTATION AND RESULT

We used Stanford parser to produce Penn Treebank. We extract the attributes of some entities by using regular expression proposed in[1]. And the sentences that don't match regular expression are passed to Rule modules that extract tuples using our proposed heuristic rules. After that search all relations and attribute to find shared attributes or relation and if found we predicts the name of supertypes using WorldNet. After that draw the result using Jgraphx in java. For example see the problem blow and its solution in (figure 3):

A medical facility has a name, address, possibly a specialty area, and the name of an administrator. inpatients have name, id number, address and phone. inpatient is treated by at least 1 or more doctors. for each outpatient we store his name, id number, address, phone, out date and method of payment.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

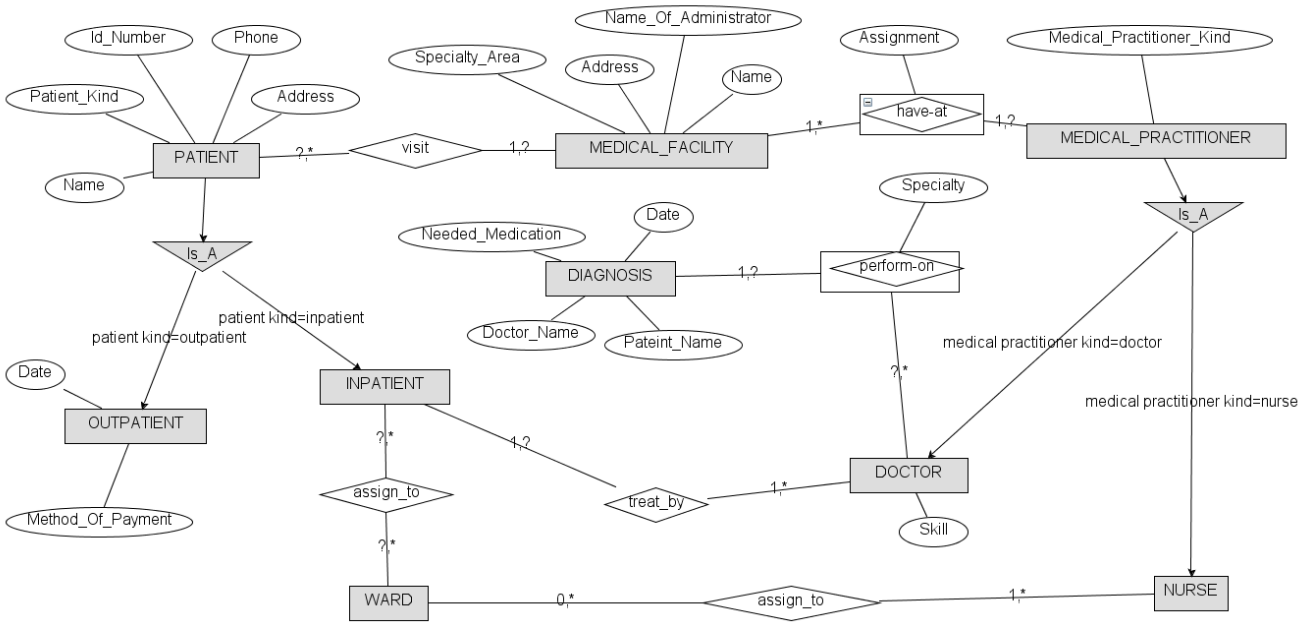


Figure 3.

inpatients and outpatients visit specific medical facility. Doctors are allowed to perform diagnosis based on their specialty. each diagnosis has doctor name, patient name, date and needed medication. Each doctor or nurse must have assignment at 1 or more medical facility. doctors have certain skills that must be recorded and accessed for a new assignment. inpatients are assigned to wards. Nurses are assigned to 0 or more wards. Each ward has at least 1 nurse assigned.

The question mark in (figure 3) represents that the min or max cardinality is not specified.

V. Conclusion and Future Works

We have described an approach for generating ERD from textual requirements using a heuristics-based system. The heuristics used are application-domain independent. We found that Penn Treebank is a promised natural language techniques because previous works used high level specific parsers elements such as Stanford typedDependencies , Link type of Link parser or others. Table [2,3] show our proposed heuristic rules and its equivalent heuristic rules proposed by Siqing Du[1] for Stanford TypedDependencies and Link types produced by Link parser. Penn Treebank are produced by many parsers (Stanford parser, openNLP, Gate parser, NLTK for python and others). We proposed rules that are decoupled from specific parsers. In the future how to use Penn Treebank in other fields such as object oriented design will be investigated.

Table 1: TypedDependencies heuristic rules equivalent to our proposed heuristic rules.

Rule No.	Stanforf TypedDependencies heuristic rules	Proposed rules
1	nsubj(w2,w1) + dobj(w2,w3)=> <w2 w1 w3>	Heuristic 1
2	nsubj(w2,w1) + prep(w2,w3) + pobj(w3,w4)=> < w2-w3 w1 w4>	
3	nsubj(w2,w1) + xcomp(w2,w3) + dobj(w3,w4)=> <w3 w1 w4>	
4	nsubj(w2,w1)+xcomp(w2,w3)+prep(w3,w4)+pobj(w4,w5)=> <w3-w4 w1 w5>	
5	nsubj(w4,w1) + aux(w4,w2) + cop(w4,w3)=> <w3 w1 w4>	
6	expl(w2,w1)+nsubj(w2,w3)+prep(w3,w4)+pobj(w4,w5)=> <w2-w4 w3 w5>	
7	prep(w4,w1)+pobj(w1,w2)+expl(w4,w3)+nsubj(w4,w5)=> <w4-w1 w5 w2>	
8	expl(w3,w1) + cop(w3,w2) + prep(w3,w4) + pobj(w4,w5)=> <w2-w4 w3 w5>	
9	prep(w5,w1) + pobj(w1,w2) + expl(w5,w3) + cop(w5,w4)=> <w4-w2 w5 w2>	
10	expl(w2,w1) + nsubj(w2,w3) + partmod(w3,w4)+prep(w4,w5) => <w4-w5 w3 w6>	
11	nsubjpass(w2,w1) + prep(w2,w3) + pobj(w3,w4) => <w2-w3 w1 w4>	
12	nsubjpass(w2,w1) + agent(w2,w3) => <w2 w3 w1>	
13	nsubjpass(w2,w1) + xcomp(w2,w3) + dobj(w3,w4) => <w3 w1 w4>	
14	nsubjpass(w2,w1) + purpcl(w2,w3) + prep(w3,w4) + pobj(w4,w5) => <w3-w4 w1 w5>	
15	nsubjpass(w2,w1) + xcomp(w2,w3) + prep(w3,w4) + pobj(w4,w5) => <w3-w4 w1 w5>	
16	rcmod(w1,w2) + nsubj(w2,w3) + dobj(w2,w4) => <w2 w1 w4>	
17	rcmod(w1,w3) + nsubjpass(w3,w2) + prep(w3,w4) + pobj(w4,w5) => < w3-w4 w1 w5>	Heuristic 2
18	partmod(w1,w2) + prep(w2,w3) + pobj(w3,w4) => <w2-w3 w1 w4>	
19	partmod(w1,w2) + dobj(w2,w3) => <w2 w1 w3>	Heuristic 3

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 2:Link types heuristic rules equivalent to proposed heuristic rules.

Rule No.	Stanforf TypedDependencies heuristic rules	Proposed rules
1	S[sp]* + O[sp]* => <S.RW S.LW O.RW>	Heuristic 1
2	S[sp]* +MV [sp]* + J[sp]* => <S.RW-MV.RW S.LW J.RW>	
3	S[sp]* + P[sp]* + J[sp]* => <S.RW-P.RW S.LW J.RW>	
4	S[spx]* + Pa +MV [sp]* + J[sp]* => <Pa.RW S.LW J.RW>	
5	S[sp]* + PP + O[sp]* => <PP.RW S.LW O.RW>	
6	S[sp]* + PP +MV [sp]* + J[sp]* => <PP.RW-MV.RW S.LW J.RW>	
7	S[sp]* + I + O[sp]* => <I.RW S.LW O.RW>	
8	S[sp]* + I +MV [sp]* + J[sp]* => <I.RW-MV.RW S.LW J.RW>	
9	S[sp]* + I[x]* + P[p]* + J[sp]* => <I.RW-P.RW S.LW J.RW>	
10	S[sp]* + I[x]* + Pa +MV [sp]* + J[sp]* => <Pa.RW S.LW J.RW>	
11	S[sp]* + I[f]* + PP + O[sp]* => <PP.RW S.LW O.RW>	
12	S[sp]* + I[f]* + PP +MV [sp]* + J[sp]* => <PP.RW-MV.RW S.LW J.RW>	
13	S[sp]* + TO + I + O[sp]* => <I.RW S.LW O.RW>	
14	S[sp]* + OF + J[sp]* => <S.RW-OF.RW S.LW J.RW>	
15	S[sp]* + TO + I[x]* + Pv +MV [sp]* + J[sp]* => <Pv.RW S.LW J.RW>	
16	SF[sp]* + O[spt]* +MV [sp]* + J[sp]* => <SF.RW-MV.RW O.RW J.RW>	
17	J[sp]* + CO + SF[sp]* + O[spt]* => <SF.RW-CO.LW O.RW J.RW>	
18	SF[sp]*+I[x]*+O[spt]*+MV [sp]*+J[sp]* => <I.RW-MV.RW O.RW J.RW>	
19	J[sp]* + CO + SF[sp]* + I[x]* + O[spt]* => <I.RW-CO.LW O.RW J.RW>	
20	SF[sp]* + O[spt]* +M[v] +MV[sp]*+J[sp]* => <M.RW-MV.RW O.RW J.RW>	
21	S[spx]* + Pv +MV [sp]* + J[sp]* => <Pv.RW-MV.RW S.LW J.RW>	
22	S[sp]* + I[x]* + Pv +MV [sp]* + J[sp]* => <Pv.RW-MV.RW S.LW J.RW>	
23	S[spx]* + Pv + TO + I + O[sp]* => <I.RW S.LW O.RW>	
24	S[sp]*+I[x]*+Pv+TO+I+MV [sp]*+J[sp]* => <I.RW-MV.RW S.LW J.RW>	
25	R + RS + O[sp]* => <RS.RW R.LW O.RW>	
26	R+RS+I[x]*+Pv[f]*+MV [sp]*+J[sp]* => <Pv.RW-MV.RW R.LW J.RW>	Heuristic 2
27	MX[spr]* + S[spxw]* + Pv +MV [sp]* + J[sp]* => <Pv.RW MX.LW J.RW>	
28	Mv +MV [sp]* + J[sp]* => <Mv.RW-MV.RW Mv.LW J.RW>	

REFERENCES.

[1] Du, S., "On the Use of Natural Language Processing for Automated Conceptual Data Modeling", Ph.D Dissertation, University of Pittsburgh, (2008).

[2] Evelyn J. Barry, Chris F. Kemerer and Sandra A. Slaughter, "How software process automation affects software evolution: a longitudinal empirical analysis ", Journal of Software Maintenance and Evolution: Research and Practice - SMR , vol. 19, no. 1, pp. 1-31, 2007.

[3] Y. G. Kim and S. T. March. "Comparing data modeling formalisms.", Communications of the ACM, Vol. 38, No. 6, pp. 103-115, 1995

[4] William J. Black. "Acquisition of conceptual data models from natural language descriptions" In 3rd Conf. of the European chapter of ACM, Denmark, 1987.

[5] G. S. Anandha Mala and G. V. Uma, "Automatic Construction of Object Oriented Design Models [UML Diagrams] from Natural Language Requirements Specification"

[6] Peter Chen, "English sentence structure and entity relationship diagrams", Information Science Vol.1, No. 1, Elsevier 127-149, 1983.

[7] Nazlia Omar, P. Hanna, and P. McKevitt, "Heuristics-based entity-relationship modelling through natural language processing", In Proc. of the Fifteenth Irish Conference on Artificial Intelligence and Cognitive Science (AICS-04), Lorraine McGinty and Brian Crean (Eds.), 302-313, 2004.

[8] Siqing Du and D.P. Metzler. "An automated multi-component approach to extract entity relationship from database requirement specification document" In 11th International Conference on Applications of Natural Language to Information Systems, Austria, 2006.

[9] Ornsiri Thonggoom (2011). "Semi Automatic Conceptual Data Modeling Using Entity and Relationship Instance Repositories" Ph.D Dissertion, Drexel University.

[10] Christian Kop, Gunther Fliedl and C. Mayr "From Natural Language Requirements to a Conceptual Model" (2010).

[11] Paulraj Ponniah "DATA MODELING FUNDAMENTALS: a practical guide for IT professionals" Wiley-Interscience (2007).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Towards a Multi-Level architecture of distributed, dynamic and cooperative workflow management

Samiha EL MESSARI, Khalid BOURAGBA*, Mohamed OUZZIF and Mounir RIFI
Laboratory RITM, CEDoc ENSEM
University Hassan II – Ain Chock
Casablanca, Morocco

***Abstract*— Workflow technology, designed to automate business processes and provide support to their management is now an active area of research. It is within this context that our work covers both the modeling workflow process and its cooperation through a system of distributed workflow management. The emergence of what might be called Cloud Virtual Enterprise , covering all partners in order to achieve an overall business process and complement the expertise of each, has created the need for the management of this exchange and for the coordination and interoperability ensuring greater autonomy and flexibility of users, within the framework of an inter-organizational workflow. In this paper we first present a literature review of workflow management and the different existing approaches about cooperation. Then, we suggest a multi-leveled architecture of a collaborative workflow management system and distributed dynamic of a virtual business on Cloud. Finally, we offer an algorithm for cooperation between the so-called private and public components workflow in the environment of a dynamic virtual enterprise. This algorithm of cooperation is divided into four phases, namely: publication, search, filter and connection.**

Mots-clés : Workflow, WFMS, Multi-level Architecture Cooperation, Virtual Enterprise, CoopFlow, cloud.

I. INTRODUCTION

WFMS (Workflow Management Systems) are systems that manage, control and support processes such as workflow. A workflow is a set of tasks that must be performed in a well-defined order by various users following specific roles by using different resources. During the execution of a workflow, different agents inter-communicate. The notion of a collaborative process or workflow is defined by the automated execution of a set of tasks required to run a business process such as booking an air ticket or planning leave of absences for staff within a company. During the execution of a collaborative process, documents and data are exchanged between the participants who might either simple human users or machines. A WFMS provides control and management of the implementation of a collaborative process based on the specification of workflow that defines the sequence of operations required to complete the business process associated with the collaborative process. Like any collaborative application, implementation of collaborative processes must meet stringent requirements in terms of safety, quality of service and transactional consistency. The majority of WFMS implements a centralized coordinating infrastructure to address these constraints. The performance of collaborative applications, however, are limited by the centralized coordination system since the most recent applications require significant flexibility in execution. This is why a decentralized WFMS provides a solution to the issue of flexibility.

Globalization, the availability of information systems through web platforms, competition and the expenses accrued by the management and automation of business processes, are among other factors that encourage companies today to be open to external partners and weave up business relationships that would create the so-called inter-organizational workflows (WFIO) [1].

Many research publications offer a whole range of workflow terminology and concepts, and the relationships between these workflows. The concepts defined by WFMC [2] and then reshaped by Van der Aalst and Van Hee in [3] are the most widely applied in business process management.

In this work, we have focused on the suggested architecture workflow management which consists of four levels, basically the level of modeling workflow process which is an activity that defines and analyzes business processes. This step precedes any decision or formulation. Then we discover the specification level workflow system that includes a number of temporal, behavioral and organizational aspects to ensure their modelization. Then we get to the levels of cooperation and coordination between business processes taking into consideration the constraints that occur in many levels regarding the limits of heterogeneity to guarantee an objective cooperation that meets the needs and requirements of each partner level. Finally there is the execution and implementation level of workflow that ensures communication with users and external applications using specific interfaces during the running time.

The present paper is organized as follows: The second part is dedicated to an analysis of research background to show the obvious need for a WFMS system to support the definition, execution, recording and control processes in the environment of the information system of an organization. The third section provides an update on interoperability problems between the workflows and the existing approaches to cooperation in a virtual enterprise context. Section IV discusses our proposed

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

multi-workflow architecture that is based on four levels of management to ensure effective analysis and verification of correct and complete workflow process. Then in the fifth section, we explain the proposed environment in a dynamic virtual enterprise cooperation algorithm. Finally, we close this work with conclusions and perspectives.

II. Preliminaries and Technical Background

A. Workflow technology

Modeling through processes, whose areas of application were relatively small in the beginning, has grown steadily. It is necessary today as an indispensable tool, for the management of organizations. The failure of organizational processes in software architectures gave, among others, birth to workflow to meet a variety of requirements including the optimization of business processes, support and control.

A workflow [4] consists of the automation of a business process, either wholly or partially, whereby documents, information or tasks are passed from one participant to another for action according to a set of procedural rules [5]. Having the workflow definition, a large number of business activities in an organization can be classified in the category of workflow. For example, placing a purchasing order, applying for a bank loan, renting a car, etc.

B. Classification of workflow systems

A major classification scheme of workflow systems has been proposed in literature review [6, 7]. McReady ranked the workflows into:

- **Ad hoc workflows** used to manage core processes in a company.
- **Collaborative Workflows** managing processes that evolves quite frequently.
- **Administrative workflows** corresponding to state-oriented processes with a well-defined procedure.
- **Workflows production** involving highly structured process with almost no changes.

C. WfMC Reference Model

The WfMC (Workflow Management Coalition) has developed a reference model for workflow technology [2]. Figure 2 illustrates this model [8].

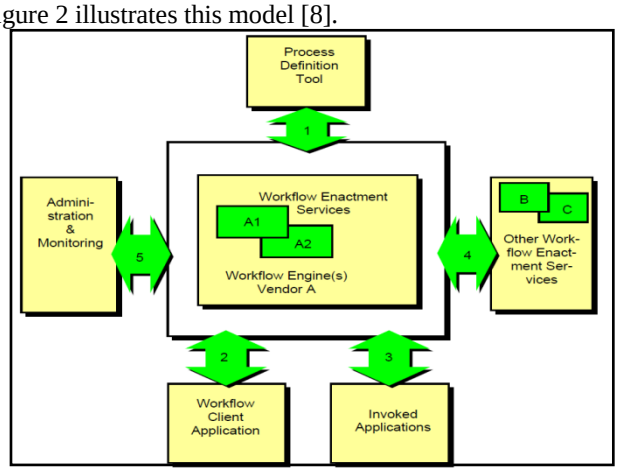


Figure 1. WfMC Reference Model.

The Workflow Management Coalition reference model, below, defines five components of workflow, the five interfaces to the workflow enactment services.

The main objective of the reference model is to provide standard interfaces and data exchange, defining specifications for interoperability between heterogeneous WF engines. It consists of a general description of the structure of a WfMS wherein five components of workflow are introduced: **Interface 1** is the process definition; **Interface 2** is the client application; **Interface 3** is the interface to the programs invoked by the business process; **Interface 4** allows one workflow system to interact with another workflow system; **Interface 5** is for the administration and monitoring of the system.

III. Interoperability and Cooperation between Workflow

In order to improve their productivity, companies expressed a great need for openness and cooperation worldwide. They need to ally with other companies with complementary skills to cooperate and carry out projects that are not within reach of a single company (eg. mergers, extension of companies into international structures, intensive outsourcing of services, etc.).

Accordingly, companies express their needs in terms of business relations via the Web and tend to automate their interaction and cooperation. This business cooperation and the automation process of manual interaction via the Web are often referred to as Business-to-Business (B2B). B2B applications refer to using computer systems (eg Web servers, network services, databases, etc.) to conduct business interactions (eg exchange of business documents, product sales, etc.) between different partners.

This has favored the emergence of new corporate structures, either on the same or on different Clouds, called virtual enterprises. These allow to associate a group of partners, scattered in time and/or space to a specific project. They combine their respective resources and expertise, and cooperate to achieve common objectives based on information technology.

A. Entreprise virtuelle

We distinguish two types of virtual enterprises; Static virtual enterprises and dynamic virtual enterprises [9]. In this context, we carried out a comparative study between the two types of businesses that have been summarized in the following table (Table 1).

The cooperation should ensure flexibility, respect of the expertise of companies, the preservation of their workflow and the integration of their existing WFMS. Subsequently, we can offer a synthesis of the main issues of interoperability of workflow (WF) in two major aspects:

1. heterogeneity between partners on different levels (part of dynamic virtual enterprise):
   - Syntactic heterogeneity (WF specification languages).
   - Semantic heterogeneity (concepts of business process modeling).

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- Contextual heterogeneity (the business environment of every enterprise).
- Technological heterogeneity (physical environment).

2. Openness and trust between partners, within organizational environment that is specific to each according to rules, data exchanged, and requirements in terms of confidentiality (under static virtual enterprise).

TABLE I. COMPARISON OF A STATIC AND DYNAMIC VIRTUAL ENTERPRISES.

VIRTUAL ENTERPRISE	
STATIC	**DYNAMIC**
-KNOWN PARTNERS / PREDETERMINED: -WF COMMON SHARED -SAFE COUPLING WF / PARTNERS (PRESET INTERACTIONS) ⇓ **NO FLEXIBLE** **NO EVOLUTIF** **CHANGE COST**	-PARTNERS NO PREDETERMINED / FREE RELATIONSHIP -WF INTERCONNECTION ON DEMAND -LOW COUPLING WF / INTERACTIONS AS NEEDED ⇓ **FLEXIBILITY** **SCALABLE STRUCTURE** **PARTNERS NOT DEPENDENT**

B. Approach CoopFlow

There are several approaches to cooperation between WF, for example: The CrossFlow [10] approach, the WISE project [11], the e-Flow [12] approach. But there is no approach to meet all the needs of cooperation that we have previously identified. We focus in this work on CoopFlow [13] approach, as it is a bottom-up approach that meets all the requirements of cooperative inter-WF.

CoopFlow is inspired by the architecture oriented services which is basically founded on three operations: publication, research and connection. Similarly, the approach consists of three steps: (1) the publication of parts of WF enterprise can be operated by other companies, (2) WF interconnection and (3) cooperation and supervision in accordance with a WF set of policy cooperation (constraints interactions). The CoopFlow approach focuses on occasional and short-term cooperation between companies where no structural constraint on cooperation is a priori defined and where partners are dynamically identified as necessary for cooperation needs. Its purpose is to allow WF running in different companies dispersed in space and/or time, heterogeneous, autonomous interacting and cooperating together.

The cooperation of enterprises can be examined under two scenarios. In the first scenario, there is a degree of trust between the involved partners who can cooperate and exchange data directly without the intervention of mediating entity. In the second scenario, there is not enough trust between partners and in this case, a trusted third party must be put in place to ensure the exchange of data as well as control and monitoring of cooperation between partners.

To achieve this purpose, we have developed a multi-layered architecture that addresses all aspects of WF. In the cooperative aspect, we have suggested an approach inspired from CoopFlow with four main steps of cooperation, namely publication, discovery, filtering and connection. To satisfy the constraint of trust, we suggest two levels of cooperation: a first level with private directory in which there is cooperation between partners with a certain degree of mutual trust; and a second level with a public directory to interconnect and to cooperate partners with less sufficient confidence.

IV. MULTI-LEVEL ARCHITECTURE FOR WORKFLOW MANAGEMENT

We offer a multi-layered distributed architecture to specify, verify, and connect to cooperate partners. This management architecture WF consists of four levels (Figure 3), namely the WF level modeling process which is an activity that involves the definition and analysis of business processes. This is the step that precedes any decision or formulation. Several modeling tools can be used to describe the behavior of workflows. We quote the following:

- Petri nets (PN) [14]: They are a major formalism for modeling WF process [15]. One of the strengths of the PN is the strong mathematical foundation and the visual representation they provide.
- UML: The state / transition diagrams are another major formalism for modeling WF process. Weissenfels and al. [16] have invested in the use of state / transition diagrams to model WF process (Mentor WfMS project).
- Finite state automaton: An automaton is a well known in the field of formal specification of model systems. This model can be used for modeling WF via labels; it allows modeling the execution of an action.

As far as the specification level of our approach is considered, we take it for granted that in our WF system a number of aspects must be specified namely, the functional aspect, the behavioral aspect, the temporal aspect and the organizational aspect.

The functional aspect is the identification of the process activities to be modeled. It is important to understand that this is not only to identify the functions of the various departments of an organization but also to distinguish the activities making the process. Functional modeling must also establish the hierarchy of activities (express possible decompositions in terms of sub-processes). Finally, the functional model must also represent the data flow associated with the activities and interdependencies of data between activities (data flow).

The behavioral aspect is an essential feature of WF since it corresponds to the dynamics of the process. The behavior is expressed by modeling flow control between activities. This

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

allows specifying the time span of the implementation of activities, their stream (sequential or parallel), the points of synchronization between activities or, contrastingly, disjunction points. In addition, the behavioral model must represent the events which trigger activities. We stress the importance of this model that allows the execution of WF.

The temporal aspect is very important in the sense that the WF template should be able to capture different aspects of business processes including structural properties, data and resources, but also temporal properties. Modeling time in the WF has been little addressed in the context of research projects. Marjanovic and Orlowska announced that three temporal constraints can be specified [17]: (1) time constraint that models the expected activity in a WF duration (time on a single or on a range of values), (2) a deadline constraint can be specified in terms of absolute time limit when an activity must begin or end during the execution of workflow, (3) an interdependent temporal constraint that determines when an activity must start / finish as to the beginning / end of another activity.

The organizational part concerns the description of the organization of corporate intervening parties. The organizational model can accurately be reflecting the organizational structure of the company, ie the hierarchical decomposition of the latter into departments and services or describe organizational units which are identified as actors. According to the method chosen, the description is more or less detailed and establishes reporting relationships between the actors and the relationships between organizational units or departments. However, whatever method is chosen, the description of the roles associated with different activities remains invariant. Roles create the interface between the organizational model and the models representing activities.

Several WF specification languages exist. Here is an example:

- BPEL4WS (Business Process Execution Language for Web Services) [18] is a modeling language based on XML WF. It defines the interactions within and among organizations and creates a multitude of Web services, either synchronous or not, in a stream of WF.
- YAWL (Yet Another Workflow Language) [19] is based on the PN with additional features to facilitate modeling of complex WF.
- WF-XML: The objective of Wf-XML [20] is to provide an interface for interoperability between management systems WF and enable companies to communicate together by their automated systems WF.

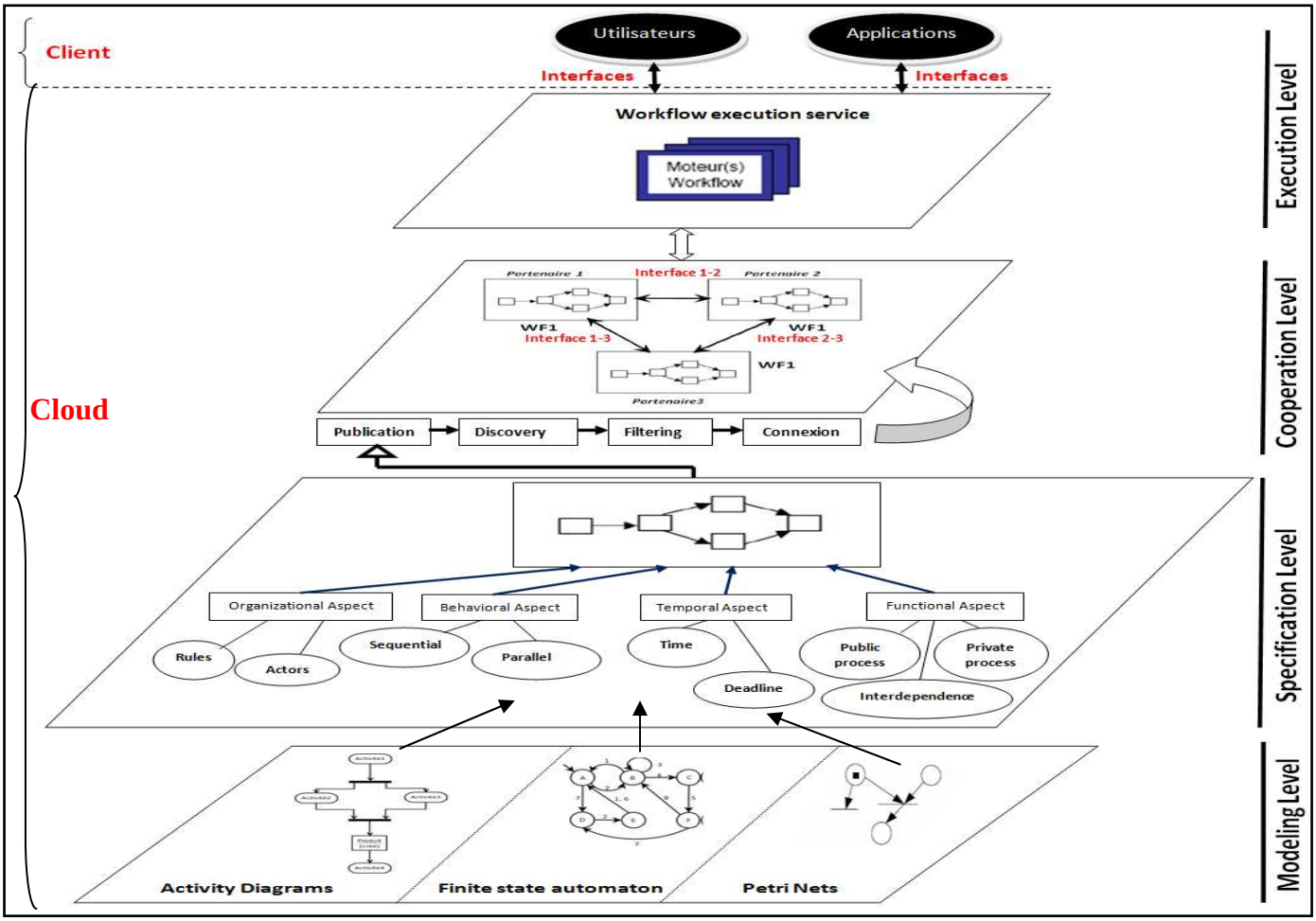


Figure 2. Multi-layer architecture for workflow management systems.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Then we get to cooperation and coordination between business processes with constraints on many levels. These constraints may be related to the limits of heterogeneity that need to be taken into consideration to ensure an objective cooperation in order to meet the needs and requirements of each partner. In this context, we have adopted a new approach based on CoopFlow which can be divided into four main stages: the stage of WF publication which in turn contains the transformation phase, abstraction and the annotation of publication of WF. The second stage concerns the discovery of WF. Then, the step of filtering the WF in accordance to criteria specific to quality service. Finally, the stage of inter-company deployment within centralized or distributed architecture.

Finally there is the execution level and the implementation of WF which ensures communication with users and external applications using very specific interfaces during running time.

To better understand the multi-layered architecture shown in Figure 2, we give an example of a WF reservation for business trip. Intuitively composed of several WF, from voyage data, airline schedules must be consulted (WF1), the flight can be booked (WF2), the hotel (WF3) and a car should go through reservation process also (WF4). These four WF can exist on the same Cloud or on different Clouds as static or dynamic virtual enterprises. The end user communicates with the system via WF client applications that are related to the runtime WF via the incorporation of interfaces that are provided by a series of calls to the API (Application Programming Interface Workflow WAPI). Then what needs to be checked is that if the same engine can perform various WF in parallel or sequentially. Otherwise, there will be a call for cooperation level to ensure the composition of WF components according to the algorithm defined in Section V. In order to meet the needs of WF users or partners, the timing of the execution of activities, their stream (sequential or parallel), the points of synchronization between activities or on the contrary and points of disjunction can be modified. This can be done by acting on the WF specification layer, which in turn calls for modeling level wherein tasks and the conditions for their implementation are defined using Petri Nets.

V. COOPERATION APPROACH BETWEEN WORKFLOW

Our approach is based on the cooperation Coopflow approach with a vision of dynamic virtual enterprises. In fact, these companies are more flexible because their structures are dynamic and depend on the needs of partners; the relationships between partners are defined at the interconnection of business.

Accordingly, we can define the relationship of an overall workflow as follows:

WF global = n WF locals + 1 coordination scheme

Each WF in our approach is considered as component with both required and provided interfaces. The cooperation between WF is done through the composition of WF components through these interfaces. This composition is governed by a contract whereby the types of documents and their used content are negotiated as well as the duration and quality of service required by the workflow of service requester.

In our approach, there are two types of WF components: private and public. We consider WF private components that are part of the same company or the same Cloud, in which the whole internal structure WF of business will be visible to all participants. This implies that there is a degree of trust and visibility between the partners (Figure 3).

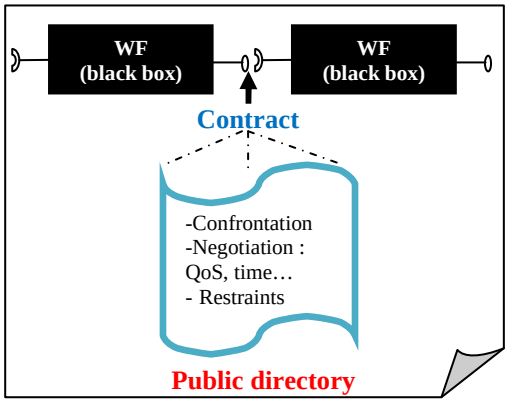


Figure 3. Publication of WF components in the public directory.

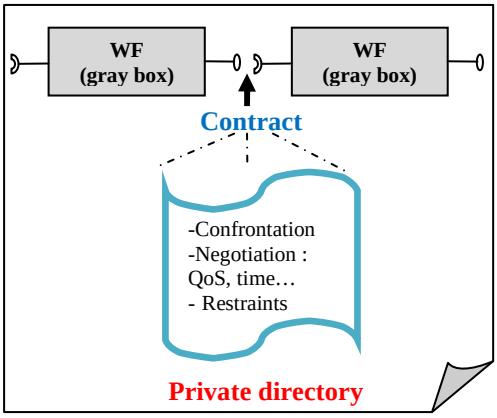


Figure 4. Publication of WF components in the private directory.

The WF Public components are represented by black dots in which the whole structure of WF must be hidden and only the necessary part to cooperation is published. We adopt the concept of black box in which communication between partners is of request / response type. This type of components is governed by short-term contracts (Figure 4).

In the publication phase all partners publish offers and requests WF service in a common directory that we have divided into two parts, namely a private directory for private WF components (gray box), and a public directory where WF public components are registered (black box). This is to record all WF activities as well as its control flow and data in these directories.

The discovery phase is triggered when a company wishes to cooperate with other companies. It begins by identifying partners with complementary skills. This is a set of cooperation policies describing the responsibilities and roles of each partner during the cooperation and coordination of their WF. The construction of the virtual enterprise is carried out through a cooperative interface matching procedure that takes into

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

account not only flow control and data but also semantic descriptions of cooperative activities.

Finally the step of connection and selection of partners can be planned into three phases, namely:

- Filtering according to specific criteria (QoS Availability ...).
- Negotiation between partners: why we need to develop protocols to reach an agreement and establish a contract.
- Distributed execution and ongoing contracts.

We propose an algorithm for the management of cooperation and networking between WF components based on the concept of publication on private and public phone directory, as well as filtering the appropriate workflow based on the criterion of better quality of service (QoS).

Algorithm : Cooperation between WF components

Implements :
Coop_WF ;

Uses :
Private_Component; Public_Component ;
Private_Directory ; Public_Direcory
%NB_WF%
%QoS%

```
Upon event <Init> do
Private_Component:=1 ;
Public_Component:=1;

Upon event <CooperationRequest to Private_Directory |
IDCoop_WF, ProcessType, ProcessData, condition> do
    For (int Private_Component =1 ; Private_Component  <=
          NB_WF ; Private_Component ++)
    {
      if ( QoS_ Private_Component = Best(QoS_
        Private_Component (NB_WF) )

        Trigger <ConnectionRequest (Private_Component) |
          IDCoop_WF, ProcessType, ProcessData, condition >;
    }

Upon event <ConnectionConfirm | IDCoop_WF> do

    Trigger <ConnectionStatus | IDCoop_WF, Ok>;
Exit ()

else
Trigger < CooperationRequest to Public_Directory >;

Upon event < CooperationRequest to Public_Directory  |
IDCoop_WF, ProcessType, ProcessData, condition > do

For (int Public_Component =1 ; Public_Component  <=
          NB_WF ; Public_Component ++)
    {
      if ( QoS_ Private_Component = Best(QoS_
        Private_Component (NB_WF) )

        Trigger <ConnectionRequest (Public_Component) |
          IDCoop_WF, ProcessType, ProcessData, condition >;
}
Upon event <ConnectionConfirm | IDCoop_WF> do
Exit ()

else
Trigger <SystemAlarm>;

Upon event <SystemAlarm> do

    Trigger <Send  Error _message "Service non-available">
```

VI. CONCLUSION AND PERSPECTIVE

Today, the automation of business processes is an important challenge. Before any development and implementation of an information system, it is essential to master the sequences of operations of these processes by automating the execution and coordination of these activities in the form of workflow. In this paper, we have initially suggested i a multi-level approach to workflow management starting from the process modeling phase to the execution phase. In a second step, we have developed an algorithm that describes a new approach to cooperation which relies on four phases, namely publishing, searching, filtering and connection, while focusing on the notion of private and public directories. However, for future research undertaking, we plan check the consistency of the execution of workflow process using the structural theory of Petri nets. We also aim to carry out our proposal by implementing it in the Bonita workflow platform taking into account the already suggested multi-layered approach and the algorithm of developed cooperation.

REFERENCES

[1] E. Andonoff, L Bouzguenda, C. Hanachi, C. Sibertin-Blanc. “Finding partners in the coordination of loose inter-organisationnal workflow”. IRIT-UTI,Université Toulouse 1, 2006.

[2] WFMC. Workflow management coalition terminology and glossary. Technical Report WFMC-TC-1011, Workflow Management Coalition, Brussels, 1999.

[3] W.M.P.van der Aalst et K. van Hee. Workflow Management : Models, Methods, and Systems. The MIT Press, ISBN 0-262-01189-1, 2002.

[4] M. Zur Muehlen. Workow-Based Process Controlling : Foundation, Design, and application of Workflow - driven Process information system. 2002.

[5] F.L. Tiplea et D.C. Marinescu. Structural soundness of workflow nets is decidable. Information Processing Letters 96, pp. 54-58, 2005.

[6] S. Bassil. Workflow Technology for Complex Socio-Technical Systems. Thèse de Doctorat, Université de Montréal, 2004.

[7] S. McReady. There is more than one kind of Workflow software. Computerworld, Vol.2, pp. 86-90, 1992.

[8] G. Zacharewicz. Un environnement G-DEVS/HLA : Application à la modélisation et simulation distribuée de workflow. Thèse de doctorat de l'université Paul Cesanne Aix-marseille III, 2006.

[9] B. T. IT-Infrastructure for Dynamic Virtual Enterprises. Master Thesis, January 2004.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[10] P. Grefen, K. Aberer, Y. Hoffer, and H. Ludwig. Crossflow : Cross-organizational workflow management for service outsourcing in dynamic virtual enterprises. IEEE Data Engineering Bulletin, 24(1) :52–57, 2001.

[11] G. Alonso, U. Fiedler, C. Hagen, A. Lazcano, H. Schuldt, and N. Weiler. Wise : Business to business e-commerce. In International Workshop on Research Issues on Data Engineering : Information Technology for virtual Enterprises, Sydney, Australia, March 1999.

[12] F. Casati and M. Shan. Dynamic and adaptive composition of e-services. Information Systems, 26(3) :143–163, 2001.

[13] I. Chebbi, S. Dustdar, and S. Tata. The view-based approach to dynamic interorganizational workflow cooperation. Data and Knowledge Engineering Journal,56(2) :139–173, 2006.

[14] Petri Nets World. Online Services for the International Petri Nets Community. www.daimi.au.dk/PetriNets, 2004.

[15] S. Jablonski et C. Bussler. Workflow Management - Modeling Concepts, Architecture and Implementation. International Thompson Computer Press, ISBN 1850322228, 1996.

[16] J. Weissenfels, P. Muth et G. Weikum. Flexible Worklist Management in a Light-Weight Workflow Management System. Proceedings of the Workshop on Workflow Management Systems at the Sixth International Conference on Extending Database Technology (EDBT'98), pp. 29-38, 1998.

[17] O. Marjanovic et M.E. Orlowska. On Modeling and Verication of Temporal Constraints in Production Workflows. Knowledge and Information Systems, Vol. 1(2), pp. 157-192, 1999.

[18] T. Andrews, F. Curbera, H. Dholakia, Y. Goland, J. Klein, F. Leymann, K. Liu, D. Roller, D. Smith, S. Thatte, I. Trickovic, and S.Weerawarana. Business process execution language for web services, version 1.1, 2003.

[19] W. M. P. van der Aalst and A. H. M. ter Hofstede. YAWL : Yet another workflow language. QUT Technical report FIT-TR-2003-04, Queensland University of Technology, Brisbane, 2002.

[20] D. Hollingsworth. An xml based architecture for collaborative process management. In L. Fischer, editor, Workflow Handbook 2002, pages 95–116. Future Strategies, Lighthouse Point (FL), 2002.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Text Summarization Using Tense Recognizing Identifier

Rajesh Wadhvani
Computer Science Department
National Institute of Technology
Bhopal, India
Email: wadhvani_rajesh@rediffmail.com

Devshri Roy
Computer Science Department
National Institute of Technology
Bhopal, India
Email: devshriroy@manit.ac.in

***Abstract*—Text summarization method finds important data and select relevant sentences to form a summary from the original text data. In this paper a new approach for extracting summary from a document is proposed. The updated information about the document is obtained in two phases. In the first phase, parsing is performed on tagged sentences governed by some rules and different chunks obtained are separated. In Second phase, summary of future chunks is found based on some parameters like word frequency, sentence position etc. To extract the future sense of sentences from the complete text, modal verbs are used. The new summary produced can now meets the need of user by making selection easier and saving reading time. We generated the summaries of documents using our approach and compared the results with respective ideal summaries that showed that sentences in the generated summaries are really helpful in extracting the important data according to user.**

Keywords: Knowledge Extraction; Text Summarization; Extract Summarization; Abstract Summarization; Part-of-speech(POS) Tagging; Scoring.

I. INTRODUCTION

vandu With the expeditious growth of technology enormous amount of data is added on WWW constantly. However huge amount of data which is available online is in different forms. Media type's learners find it intensely onerous to extract useful information from the given data. Internet carries an extensive range of information resources and when user present queries for it, it gives result in just within seconds. Although the results obtained may or may not be relevant and reliable text to satisfy user. Generally user reads the complete document to decide its favourability which is a time consuming and complex process. To make judgement whether the document is useful or not, user needs some succinct support. User needs concise overview of large sets of article in very less time. Regular way to achieve this is to apply Information retrieval techniques. Information retrieval does searching by keeping highly relevant document on top. However the result obtained still cannot be managed by user. Thus techniques are needed that performs fast text processing and give efficient outcome within time. Test summarization is such technique; it is a process of distilling the most important information from a source for a particular user. It serves as a solution for this issue. Now to extract the informative data from the original document different types of summarization techniques are used we focuses on extract based summarization technique [1].

Architecture is proposed that finds the summary by recognizing tense of different sentences present in the document. Each document gives knowledge about certain topic and for explaining that it sometimes considers past work results while other times it tells about future events. Thus if the user is interested only in future work of a topic and want to track the updated information over time, it can be extracted by considering future sense in the document. In this method user can generate information in his choice of sense i.e. Present, Past and Future information about topic from the document. No doubt present happening events and real time information is more important than other parts in the text. Thus here one more advantage is provided to the user to know about the events that will happen in future. For example if user is reading about conference, he/she can easily find about the upcoming conferences dates and places. After reading reports he can decide what the areas on which he should work. Updated document generated form summarizer will contribute towards acquisition of new informative knowledge updates. We explicate the conventional sentence selection procedure by tagging and appraising tense aspects in sentences for parsing. Using parse trees future related chunks are separated to be used in final summary. In this paper also the most relevant updated information will be presented at the top of summarized document. 'Term Frequency' based method is used for scoring of sentences to obtain the summary in which higher scored sentences are at the top.

The organization of remaining sections in this paper is as follows: Section II outlines the classification of various existing summarization and illustrate about the related work in extraction based summarization. In section III the proposed algorithm is discussed in detail. Section IV gives detailed evaluation procedure. Finally the paper is concluded with future research directions in section V.

II. RELATED WORK

Text summarization has been widely studied in recent years. Radev et al. (2002) defined a summary as a, "text that is composed from one or more texts that fetch useful from the original text and then the resultant text is no longer than half the original text". To identify the most important content out of rest of the text the approaches are of two types: Extractive summarization and Abstractive summarization [2, 3]. In Extractive summarization approaches, the aim is determining most crucial concepts from the input document and giving

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

affiliated sentences recover in the document as an output. In contrast Abstract summarization approach is very similar to human summarization which uses linguistic methods to find new sentences, concepts and topics that best describes the original text document. Since last few years some systems have been developed that generate abstract of a document using the latest natural language processing tools [4]. These systems extract phrases and lexical chains from the documents and use generative tools to produce abstract of document. Gong et al. [5] proposed a method using latent semantic analysis(LSA) and relevant measures to summarize matter depending on cosine similarity measure. Text information can be analysed by structuring the text using different grammar rules and additional constraints 6].

WorldNet is a lexical reference system based on psycholinguistic concept of human lexical memory, it coordinate lexical information in terms of word meaning and makes a network of words linked by lexical and semantic relations. Summarization approaches are also based on statistical and linguistic features like in [7] where sentences are ranked for potential inclusion in the summary using a weighted combination function of sentence weight, subject weight. Some techniques attempts to find important word and phrases in the text using which users can browse hierarchy for extracting knowledge, refinement of query and personalization of search results [8]. In [1] different text summarization Extractive techniques are explained some of which uses linguistic methods to find new concepts and expression in the text. While in [9] the authors showed the impact of Lexical chain scoring methods on text summarization. Words in sentence are syntagmatically and paradigmatically related thus assurance of the correct sense of the word is important. Assigning sense to a content or function- word in corpus is problematic in tagging.

Different schemes have been proposed making use of statistical, linguistic, support vector machines or symbolic learning knowledge that can be categorized into: morphological analysis and part-of-speech (POS) tagging based. The part of speech tagging is a important step for knowledge extraction on the summarization method of an actual system [10]. Morphological analysis is performed to add different grammar rules and additional constraints that quantifies common and uncommon phenomenon on lexical basis [11]. J. Gimenez and L. Marquez presented the SVM Tool, a simple, flexible, effective and efficient partofspeech tagger based on Support Vector Machines [12]. And it is also found that when symbolic learning knowledge is effective when applied to POS tagging [13]. The classical approaches of tagging use standardize tag sets such as Penn/Brown set. Part-of-speech tagging by latent analogy is competitive with Markovian technique, for more accuracy it can be improved based on emerging principle of Latent analogy [14]. The structure of natural language is enhanced after POS tagging, because there exists relationship among words. Dependency parsers can extract these types of relationships between words of sentences. Leonardo Lesmo in [15] found that semantic information can be described in Syntactic Dependency Parser.

III. PROPOSED METHODOLOGY

We propose a model which is useful for text summarization of documents by using tense recognition of sentences. Applying this model user can choose type of summary he is interested in. After tenses recognition we will make separate chunks of past, future and present. If reader wants to obtain summary which should only contain past information he can check on past summary. In similar way if user want to know only about upcoming information he can select option of future based summary as we know a future tense is a verb form that mark the events that have not happened yet, but expected to happen in future. Here summary is generated to find the upcoming events based on future tense sentence.

The architecture and implementation of the proposed system for retrieving information from a document is concerned with the methods of summarization. The main processing in our system is done for English language only. Initially the text document is given as the input and then following procedure is followed to find the summary:-

i) Sentence segmentation and Part-of-speech tagging
ii) Sub sentence Parsing
iii) Grouping chunks based on tense
iv) Sentence Scoring

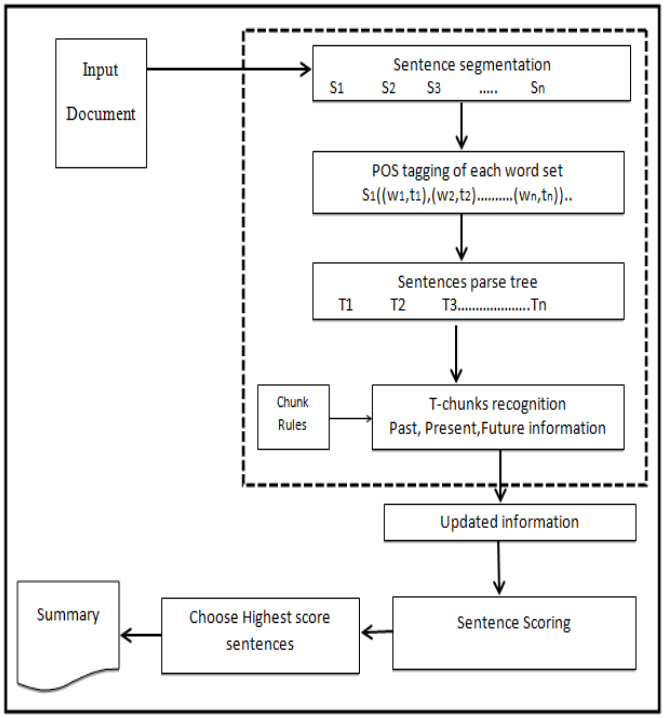


Fig. 1: Flowchart of proposed methodology

A (i) Sentence Segmentation

By splitting the colossal text (which contains all things) into sentences, we have illuminative data. Irrelevant or sentences which give metonymy meaning can be ignored. It avoids unnecessary processing of all the repetitive text in the document. Each document D_m consist of n sentences named as $S_1, S_2, ..., S_n$ where each sentence S_n is collection of w text subsets $w_1, w_2, ..., w_n$.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Sentence Segmentation Algorithm -

i) Read the text in the given file into a String
ii) String paragraph = new Scanner(new File ("doc.txt")).useDelimiter("Z").next();
Here Z means the end of file
iii) Split the string into sentences by providing the line separators
String[ ] result=paragraph.split("(. || ? || !)*(_)(. || ? || !)*");
iv) Each sentence is stored in an arrayList result [ ] for further use
v) End

(ii) Sentence POS Tagging

This is speech identification phase. Part of speech tags give concerning information about the role of a word in its lexicon. It may also provide information about the verbalization of a word. POS tags are a valuable part of a language processing tasks; they provide favourable information to other components such as a parser or a named-entity recognizer. Most of the POS tagging is done by Brills transformation based learning which take unidirectional approach. Here for tagging unidirectional approach is used. New tag (t_0) is figured considering the previous tag (t_{-1}), for finding tag (t_1), tag (t_0) is considered; this is done in backward direction. Each sentence POS tagging is performed using Stanford POS tagger [16].

For sequences of word in sentence $w_1, w_2, w_3, ..., w_n$ a sequence of tags $t_1, t_2, t_3, ..., t_n$ are assigned respectively based on Peen Treebank tag dictionary.

Event Sub sentence is generated with word w_i and its corresponding tag t_i

$S[(w_1, t_1), (w_2, t_2)....(w_i, t_i)]$

POS Tagging Algorithm-

i) Initialize an object of MaxentTagger giving the corpus to be used as the argument.
ii) Read the file to be tagged.
iii) Create a new file called out.txt to store the tagged text.
iv) Split the text read into paragraphs. para=untagged.split();
v) Loop through all paragraphs. for i=1 to n do tagged=tagString(para[i]) out.append(tagged) end

B. Sub Sentence Parsing

Parsing works on the grammatical structure of sentences in which group of words are considered collectively. We make use of statistical parsers that rely on a corpus of training data which have already been annoted and the implementation considers the lexicalized dependency. It produces the dependent tree structure of tagged text. We know that verbs are the essential part of sentences and they can help to detect the tense of a sentence. Modal verbs such as can, could, may, might, will, would, must, shall, should, ought to are used to judge the future situation.

Different parse tree are generated from parser as $T_1, T_2, , T_n$. The head of each constituent of sentence is identified as behaviour of all the other words depends on the head. It make more sense to elect the verb as head or the root node while determining dependencies.

For example on giving this sentence as a input of parser is *"The student felt nervous before the speech* the tree structure obtained for the sentence will be like Fig 2:

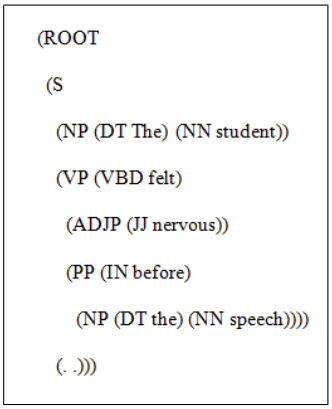


Fig. 2: Sentence parse tree structure

C. Grouping Chunks Based On Tense

Corresponding to every sentence, there exists parse tree. Now we can classify all these sentences into particular classes. Traverse each parse tree and separate present, future and past chunks. Thus complete sub sentence set can be distributed into different groups as past data, present data and future information. Each pair $< verb, setofdependants >$ is matched with knowledge about verb sub categorization in order to find out the syntactic role of each argument.

After separation of these three chunks we select only the future information set since we are interested in finding updated information (as Past information set is not emphasized on judging). Now store this future chunked information in document. The future tense is used to tell what "will" happen. E.g. *I **will** go to the beach next month. I **shall** write the letter next week.* For actions that will occur in the near future, the present tense is more commonly used.E.g. *Next year I am going to Spain.*

Thus future chunks detected by traversing to search these rules in each sentences:

- Will/Shall + base form of the verb
- Am/is/are + "going to" + base form of verb
- Will/Won't + be +verb(Present Participle)

Next step is to score sentences of future information set. At this stage the paragraphs containing the future sentences are selected. These future chunks are merged to form a document named as updated document, which will give information about work to be done in future. Now this updated document is used to score each sentence based on term frequency, position of sentence and other parameters.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

D. Sentence Scoring

Scoring can be performed by different methodologies using semantic analysis [17]. For scoring each individual sentence of updated documents a score function is defined using following parameters:

(i) Word frequency

In a document to each term a weight is assigned according to the number of occurrences of the word in that document. To prevent the bias towards longer documents word frequency is divided with the total number of words in document. Therefore, Word frequency of a word,$Wf$is calculated as:

$$Wf_i = \frac{count_i}{\sum n_k} \tag{1}$$

(ii) Sentence Score

Sentence score is the Word frequency score of a sentence Si divided by number of words in sentence.

$$SentenceScore(S_i) = \frac{\sum_i Wf_k}{k} \tag{2}$$

where K is no of words in sentence Si.

(iii) Position score

As we have seen sentence present at the starting of the text give general information of document and ending mostly concludes the document. So these starting and ending lines are important to be considered to form a summary. To give high weightage to first sentence we use formula as given below where sentence Si is at position Pi and N is the total number of sentences in the document.

$$PositionScore(S_i) = \frac{N - P_i}{N} \tag{3}$$

(iv) Future tense sentence score

We are also including another parameter for giving high weightage to sentences those are in future tense. Thus future tense score is calculated as:

$$Futurescore = \frac{\text{no. of future term within the sentence}}{\text{total no of term in a paragraph}} \tag{4}$$

In this way for a sentence Si the total weighted score function is calculated using equations (2), (3)and (4).

Total score S_i = (Sentence Score + Future Score) * Position Score

IV. EVALUATION

To evaluate summary we have develop an environment to compute and visualize the summary of the text .Document used in this paper are used from DUC 2007 corpus. D set from this document has been used from this corpus for experimental purpose. On each document this method is performed and for tagging Stanford tagger is used. Using the scoring scheme we found that most important sentences are at the top most of the time. Here we rely on the results of part-ofspeech tagger, parser and the verb filtering method. For evaluating the performance of generated future tense chunked based summary, represented as M is judged with human manually generated summary represented by H. To determine the quality of the summary precision and recall are the two performance measures. Generated summary is compared to an "ideal" reference summary to find a sentence overlaps. To construct the reference summary, a group of humans are asked to extract sentences and then the sentences chosen by majority of persons are included in the reference summary calling as H.

Precision: Precision is the fraction of the obtained sentences that are relevant to the search. It is calculated as the number of sentences common in both M and H divide with number of sentences in H.

$$Precision(p) = \frac{|M \bigwedge H|}{|H|}$$

Recall: Recall is the fraction of the sentences that are relevant to the need that are retrieved successfully. It is calculated as the no of sentences common in both M and H divide with the no of sentences in M.

$$Recall(R) = \frac{|M \bigwedge H|}{|M|}$$

F-measure can be calculated as follows:

$$F - measure = \frac{2*P*R}{P+R}$$

In Fig. 3, the horizontal line indicates the document numbers and vertical indicates the recall and precision percentage- sentences present both in the reference summary and the summary of our summarizer.

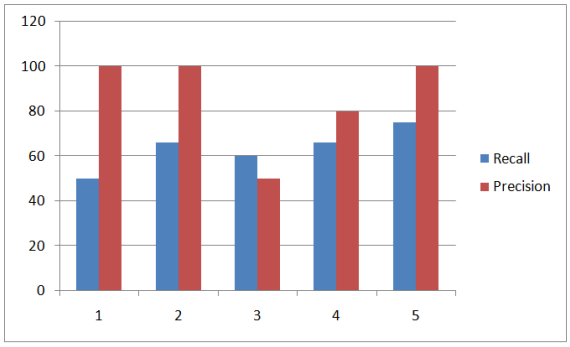


Fig. 3: Sentence overlaps between our summary and the reference summary using precision and recall

In Fig. 4, all three measures are shown for comparison of reference summaries.

V. CONCLUSION AND FUTURE WORK

We have presented an approach for determining essence of necessary sections of text which are best suited to represent the available text according to the user views and interests. Moreover, our approach provides an improvement over different views of extraction based summarization techniques. Summary generated after recognizing the tense of sentences is helpful in filtering information and thus is advantageous for information retrieval.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

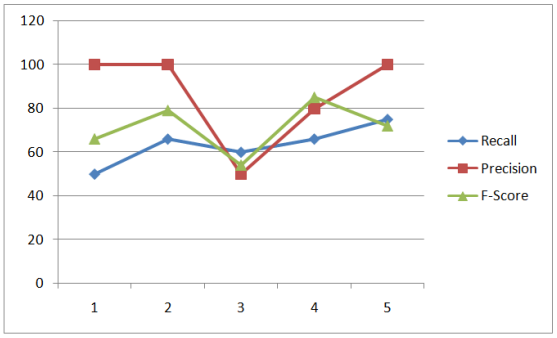


Fig. 4: Comparison of Reference summaries

The research work can be extended in many ways. We have used it for only English language; this can be also used for other language. The weighting technique can be modified giving different weights to tagged terms, even other scoring schemes can also be used. This proposed method can also be judged using different parsing methods. Another important aspect in summarization is updated summary for multi document based on query in knowledge domains using clustering, thus we can extend this method from the generic ones.

REFERENCES

[1] Vishal Gupta, Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", Journal of Emerging Technologies in web Intelligence, Vol. 2, No. 3, 2010.

[2] K. Knight and D. Marcu, "Summarization beyond sentence extraction: a probabilistic approach to sentence compression", Artifcial Intelligence, pages 91-107, 2002 Elsevier Science.

[3] H. Jing and K. McKeown, "Cut and paste based text summarization", In Proceedings of NAACL 2000.

[4] http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp_tools.html

[5] Y. Gong and X. Liu., "Generic Text summarization using relevance measure and latent semantic analysis", In Proceedings of SIGIR 2001.

[6] C. Lakshmi Devasenal and M. Hemalatha, "Automatic Text Categorization and Summarization using Rule Reduction", International Conference On Advances In Engineering, Science And Management (ICAESM -2012) March 30, 31, 2012.

[7] Rushdi Shams, M.M.A. Hashem, Afrina Hossain, Suraiya Rumana Akter, and Monika Gope, "Corpus-based Web Document Summarization using Statistical and Linguistic Approach", International Conference on Computer and Communication Engineering (ICCCE 2010)..

[8] Ferragina, P., Gulli, A., "A Personalized Search Engine Based on Web-Snippet Hierarchical Clustering", In Proceedings of WWW, ACM 1595930515/05/0005 Chiba, Japan Conference (2005).

[9] William Doran, Nicola Stokes, Joe Carthy, John Dunnion, "Assessing the Impact of Lexical Chain Scoring Methods and Sentence Extraction Schemes on Summarization", Lecture Notes in Computer Science Volume 2945, Springer 2004, pp 627-635.

[10] Ahmed Amrani, Yves Kodratoff, and Oriane Matte-Tailliez, "A Semi-automatic System for Tagging Specialized Corpora" Lecture Notes in Computer Science Volume 3056,(2004) Springer , Heidelberg, Germany , pp. 670-681.

[11] Karlson, F., Voutilainen, A., Heikkila, J., Anttila, A., "Constraint Grammar -A Language Independent System for Parsing Unrestricted Text", Mouton de Gruyter, Berlin (1995).

[12] Gimenez, J., Marquez, L., "SVMTool, "A general POS tagger generator based on Support Vector Machines", In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 4346 (2004).

[13] Roth, D., Zelenko, D., "Part-of-Speech Tagging Using a Network of Linear Seperators" In Proceedings of the 36th Annual Meeting of the ACL Coling, Mondreal, Canada (1998).

[14] Jerome R. Bellegarda, "Part-of-Speech Tagging by Latent Analogy", IEEE Journal Of Selected Topics In Signal Processing, Vol. 4, No. 6, December 2010.

[15] Leonardo Lesmo, "Use of Semantic Information in a Syntactic Dependency Parser", EVALITA 2012, LNCS 7689, pp. 1320, 2013 Springer-Verlag Berlin Heidelberg 2013.

[16] http://nlp.stanford.edu/software/tagger.shtml

[17] R.V.V Murali Krishna and Ch. Satyananda Reddy, "A sentence scoring method for extractive text summarization based on Natural language queries", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No. 1, 2012.

[18] http://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/tagger/maxent/MaxentTagger.html

[19] DUC 2006 http://www-nlpir.nist.gov/projects/duc/

AUTHORS PROFILE

Prof. Rajesh Wadhvani B.E in Computer Science from Rajiv Gandh Technical University,M.Tech in Computer Science from Maulana Azad National Institute of Technology Bhopal, Persuing PhD in Computer science from Maulana Azad National Institute of Technology Bhopal. Presently Working as Asst. Prof in Department of Information Technology in Maulana Azad National InstituteTechnology, Bhopal.

Dr. Devshri Roy Ph.D from IIT Kharagpur, Specialization in Application of Computer and Communication Technologies in E-learning , Personalized Information Retrieval , and Natural Language Processing. Presently Working as Associate Prof. in Department of Information Technology in Maulana Azad National Institute of Technology, Bhopal.

A Game Theoretic Formulation for Strategic Sensing in Cognitive Radio Networks : Equilibrium Analysis and Fully Distributed Learning

Sofia Bouferda
RITM-ESTC-GREENTIC
ENSEM-Hassan II University
Casablanca, Morocco
sofiabouferda@gmail.com

Essaid Sabir
ENSEM-GREENTIC
Hassan II University,
Casablanca, Morocco
e.sabir@ensem.ac.ma

Aawatif Hayar
ENSEM-GREENTIC
Hassan II University,
Casablanca, Morocco
a.hayar@greentic.uh2c.ma

Mounir RIFI
RITM-ESTC-GREENTIC
Hassan II University,
Casablanca, Morocco
rifi@email.com

***Abstract*—In cognitive radio (CR) systems, the licenced bands are opened to secondary users (SU), when they are not used. To support this reuse the CR users must sense the environment, and use the spectrum when its not utilized by primary users (PU). For these reasons, the spectrum sensing is an important topic in cognitive radio networks (CRN). In this paper, we consider a distributed opportunistic access (D-OSA), in which CR users attempt to access a channel licensed to a primary network. In this context, we formulate the problem of designing the equilibrium sensing time in a distributed manner, in order to maximize the throughput of CR users while guarantying a good protection to the PU. Next, we study the Nash equilibrium of the system, we also propose a combined learning algorithm for continuous actions that is fully distributed, and allows to the CR users to learn their equilibrium payoffs and their equilibrium sensing time. Finaly we give a comparison between the proposed solution and a centralized one. The simulation results show that the system can learn the sensing time and converge to a unique Nash equilibrium, which come to prove the theoretical study. The comparison between the proposed solution and the centralized one shows an expected result, which is the higher performances of the centralized method, in terms of throughput and sensing time, but as we can see in the simulation results, the difference is slight.**

Keywords— Distributed Spectrum sensing, Nash equilibrium, combined learning

I. Introduction

The great development of communication has created huge demand of radio spectrum, which has become a scarce resource. Nevertheless, it has been found that the major licensed bands are underutilized and some of the remaining bands are heavily used [1]. This is in a part due to the actual resource allocation strategies, which allocate a fixed frequency band to a single licensed system.

Cognitive radio is a promising solution to this radio spectrum wastage, it allows the simultaneous access to the same spectrum band by PU and secondary users (SU), by using spatial, temporal and frequency spectrum holes left by idles PUs [2]

Spectrum sensing has been identified as a key enabling cognitive radio to not interfere with PUs, by reliability detecting PUs signals. It motivates the concept of opportunistic spectrum access (OSA), which allows secondary cognitive radio (CR) systems to opportunistically exploit the under-utilized spectrum. A basic component of OSA is a sensing strategy at the MAC layer for spectrum opportunity tracking. Due to hardware limitations, energy constraints, a secondary user may not be able to sense all channels in the spectrum simultaneously, and due to throughput constraints Secondary users can not sense channels for a long time, in general the longer the sensing time, the less time available for the CR user to transmit data during the rest of the time frame. In this case a sensing strategy for intelligent channel selection and sensing time optimisation is necessary.

The presence of high priority primary users and the requirement that the cognitive users should not interfere with them define a medium access paradigm called cognitive medium access. Cognitive medium access control is relevant to both the cognitive radio and the coexistence community. Cognitive medium access control protocols play an important role to exploit the spectrum opportunities, manage the interference to primary users, and coordinate the spectrum access amongst secondary users [3], [4], [5]

A. Literature review

Cognitive medium access control game formulations have been studied in [8], [9], [10] using different approaches. Cognitive medium access control coupled with state estimation and detection for identifying and exploiting instantaneous spectrum opportunities is one of the central challenges. In addition, the opportunistic spectrum access needs to the take into consideration energy constraints of sensing, exploration, exploitation and transmission. Then the question is how to develop cognitive sensing and access strategies that learn from observations and offer improved fairness and performance over time [9], [10] without central controllers or dedicated communication/control channels. In [11] the authors examined evolutionary game theoretic perspective of the classical medium access control. Cognitive considerations are not discussed. [5] proposed memory management for distributed spectrum sharing under limited spectrum sensing.

In [8] the authors examine cognitive medium access and the

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

relation to competitive multi-armed bandit problem. The arm bandit problem is well understood for single CR, which wishes to opportunistically exploit the availability of resources in the spectrum. For multiple users however the outcomes are not clear. This difficulty is in particular due to the interactive behavior of the underlying processes of decision making in dynamic environment. The authors proposed a Bayesian approach to identify a trade-off metric between exploring the availability of other free channels/time slots and exploiting the opportunities identified. In [10], a game theoretic learning and pricing have been proposed. In the above references, stochastic formulations of the medium access problem are examined. These formulations often lead to intractable configurations. Authors in [19], proposed a new two-step game where sensing and opportunistic access are jointly considered. A full characterization of the Nash equilibria and analysis of the optimal pricing policy, from the network owner view, for both centralized setting and decentralized setting, are also provided. Next, a combined learning algorithm that is fully distributed and allows the cognitive users to learn their optimal payoffs and their optimal strategies in both symmetric and asymmetric cases is proposed.

B. contribution

In this paper, we propose to associate game theory to learning strategies in cognitive medium access, to find the equilibrium sensing time. To the best of our knowledge, this is the first paper devoted to analyze distributed sensing. Meanwhile, the related literature usually considers the sensing time from an optimization perspective [18] [13] [5]. In contrast to the classical literature of medium access games, which does not focus on the random nature of cognitive radios, we propose a fully distributed strategic learning to learn the equilibrium payoff and the associated equilibrium strategies. Moreover, we provide many insightful results to understand the possible relationship between sensing time and transmit probability. Next, we analyze the impact of starting point and speed of learning on convergence to Nash equilibrium. Finally we propose a comparison in terms of sensing time and throughput between the proposed solution and a centralized one.

C. Organization of the paper

This paper will be organized as follows: In section II, the system model, the main notations and spectrum sensing preliminaries are presented. In section III, we describe the utility function of the game and equilibrium analysis, in section IV, we propose a distributed learning algorithm, and in section V, a comparison between the proposed solution and a centralized one. Performance evaluation and results analysis are provided in section V.

II. System Model And Main Notations

A. System model

We consider a secondary network that coexists with a primary network where each PU is licensed to transmit whenever he/she wishes for most of time except for the case when the channel is occupied by another PU. The duration of a primary frame is denoted T. We consider that we have N SU trying to access the spectrum of the PU. Throughout this work the following consideration are taken into account:

- Energy-based spectrum sensing: the primary network activity is determined by measuring the signal strength traveling over the channel. If the received signal power exceeds some given threshold, the channel is declared busy, it is declared idle in the other case.
- Imperfect sensing in the sense that SUs may declare a busy channel while it is idle (false alarm).
- Random access for data transmission of CRs. Here, we consider that SUs follow a slotted aloha-like protocol to transmit data.

During primary user's activities, each SU i receives some given signal. SU i samples the received signal at sampling frequency f_s without loss of generality, we assume that all SUs use the same sampling frequency. The discrete received signal at the SU i can be represented as:

$$Y_i = \begin{cases} h_i.S(t) + n(t) & : \text{Hypothesis } H_1(Busy) \\ n(t) & : \text{Hypothesis } H_0(Idle) \end{cases} \quad (1)$$

Where h_i is the channel gain experienced by SU i and n(t) is an circular symmetric complex Gaussian noise with mean 0 and variance $E[|n(t)^2|] = \sigma$. The channel state is considered as the binary hypothesis test H_0 and H_1.

B. Energy based spectrum sensing

Spectrum sensing is often considered as a detection problem. Many techniques were developed in order to detect the holes in the spectrum band. Focusing on each narrow band, existing spectrum sensing techniques are widely categorized into energy detection [6] and feature detection [7]. Although this is not a restriction of our work, we will use energy detection throughout the paper.
Let τ_i be the sensing time and N the number of considered samples. Thus, we have $N = [\tau f_s]$.it follows that the average energy detected by SU i is:

$$T_i(Y) = \frac{1}{N} \sum_1^N |Y_i(t)|^2 \quad (2)$$

C. Imperfect sensing

We consider throughout this work a scenario where the spectrum sensor has imperfect detection performance. In other terms, each SU i has a false alarm probability P_{fi}, i.e., the probability that the channel is sensed to be busy while it is actually idle. Let ϵ denotes the threshold which specifies the collision tolerance bound of PUs. Then:

$$P_{fi}(\epsilon, \tau_i) = Pr(T_i(Y) > \epsilon \backslash H_0) \quad (3)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

We assume that the primary signal is i.i.d complex Phase shift keying modulated, with zero mean. We also assume that the primary signal and noise are independent. Using energy detection, the false alarm probability $P_{fi}(\epsilon, \tau_i)$ can be estimated by the following:

$$P_{fi}(\epsilon, \tau_i) = \frac{1}{2} erfc\left(\left(\frac{\epsilon}{\sigma^2} - 1\right)\sqrt{\tau_i f_s}\right) \quad (4)$$

Where the complementary error function is given by:

$$erfc(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{\left(\frac{-t^2}{2}\right)} dt (5)$$

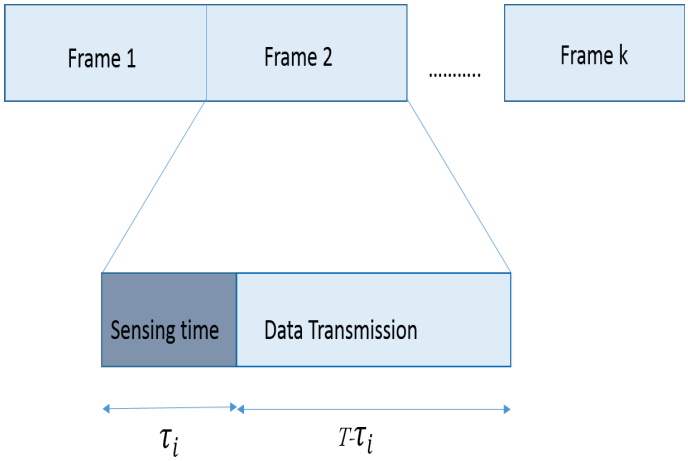


Fig. 1. Frame structure of the cognitive radio network.

Each SU i senses the channel for a duration τ_i Fig. 1. This duration affects directly the false alarm probability $P_{fi}(\epsilon, \tau_i)$ that the tagged SU i will experience. This relationship could be stated as: long sensing time leads to a more accurate sensing operation (equivalently low false probability) but generates short transmission opportunity and high competition. Thus, the experienced throughput would be affected drastically by the outcome of the sensing operation. The issue is then how to find a good trade-off between sensing and received throughput. In [13], it is shown that there exists an optimal sensing time that maximizes the throughput of CR users. Similar result can be found in [14]. In both papers, authors consider the case of centralized control, in this work unlike the papers mentioned above we consider equilibrium sensing time in a distributed framework.

III. Equilibrium analysis

In absence of primary network activity (i.e., under hypothesis H_0), the average Throughput of SU i is given by:

$$U_i(\tau_i) = \frac{T - \tau_i}{\tau_i}(1 - P_{fi}(\epsilon, \tau_i)) P_i \prod_{j \neq i} (1 - P_j (1 - P_{fj}(\epsilon, \tau_j))) \quad (6)$$

In the proposed game G each SU chooses an appropriate sensing time to maximize its utility function, which is equivalent to its throughput. In this context, it is important to ensure the stability of the system. A concept which relates to this issue is the Nash equilibrium. As definition in [15], a pure strategy profile $\{\tau_j^*\}_{j \neq i}, j \in \{1, ..M\}$ is a Nash equilibrium of the proposed game if, for every player i (i.e. SU i):

$$U_i(\tau_i^*, \tau_{-i}^*) \geq U_i(\tau_i, \tau_{-i}^*) \forall i \in \{1, ..M\} \quad (7)$$

A Nash equilibrium can be regarded as a stable solution, at which none of the users has the incentive to change its sensing time τ_i.

A. Existence of the Nash equilibrium

Proposition 1: Game G admits at least one Nash equilibrium.

Proof: The conditions for the existence of Nash equilibrium in a strategic game are given in [16]:

1) The set τ_i is a nonempty, convex, and compact subset of some Euclidean space for all i.
2) The utility function $U_i(\tau_i; \tau\_i)$ is continuous on τ and quasi-concave on τ_i.

On one hand and according to the above description of the strategy space, it is straightforward to see that it is nonempty, convex and compact. On the other hand, the utility function U_i is a continuous and quasi-concave function. Hence, game G admits at least one Nash equilibrium.

$$U_i(\tau_i, \tau_j) = \frac{T - \tau_i}{\tau_i}(1 - P_{fi}(\epsilon, \tau_i)) P_i \prod_{j \neq i} (1 - P_j (1 - P_{fj}(\epsilon, \tau_j))) = \frac{T - \tau_i}{T}\left(1 - \frac{1}{2} erfc\left((\frac{\epsilon}{\sigma^2} - 1)\sqrt{\tau_i f_s}\right) P_i \prod_{j \neq i}\left(1 - P_j\left(1 - \frac{1}{2} erfc\left((\frac{\epsilon}{\sigma^2} - 1)\sqrt{\tau_i f_s}\right)\right)\right)\right) \quad (8)$$

Let G be the primitive of $e^{-\frac{t^2}{2}}$,

$$erfc(x) = \frac{2}{\sqrt{\pi}}(G(+\infty) - G(x)) \Rightarrow erfc'(x) = \frac{-2}{\sqrt{\pi}} G'(x) = \frac{-2}{\sqrt{\pi}} e^{-\frac{x^2}{2}} \quad (9)$$

For simplicity of the derivation let put:

$$A = \frac{\epsilon}{\sigma^2} - 1. \quad (10)$$

$erfc\left((\frac{\epsilon}{\sigma^2} - 1)\right)$ is then written as follows:

$$erfc(A\sqrt{\tau_i f_s}) \quad (11)$$

From (9) we can deduce the fisrt derivative of (11) :

$$erfc'(A\sqrt{\tau_i f_s}) = \frac{A f_s}{2\sqrt{\tau_i f_s}} \frac{-2}{\sqrt{\pi}} e^{\frac{-2\tau_i f_s}{2}} = -A\sqrt{\frac{f_s}{\tau_i \pi}} e^{\frac{-A^2 \tau_i f_s}{2}} \quad (12)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\frac{\partial(U_i(\tau_i,\tau_j))}{\partial\tau_i} = K\left(\frac{-1}{T}\left(1-erfc(A\sqrt{\tau_i f_s})\right)+\frac{T-\tau_i}{T}\frac{1}{2}A\sqrt{\frac{f_s}{\tau_i\pi}}e^{\frac{-A^2\tau_i f_s}{2}}\right) \quad (13)$$

where:

$$K = P_i\prod_{\mathrm{J}\neq i}\left(1-P_j\left(1-\frac{1}{2}erfc\left((\frac{\epsilon}{\sigma^2}-1)\sqrt{\tau_i f_s}\right)\right)\right) \quad (14)$$

To simplify the second derivative let put:

$$G(\tau) = \frac{T-\tau}{T}\frac{1}{2}A\sqrt{\frac{f_s}{\tau\pi}}e^{\frac{-A^2\tau f_s}{2}} \quad (15)$$

$$\frac{\partial G}{\partial\tau} = \frac{-A}{2T}\sqrt{\frac{f_s}{\tau\pi}}e^{\frac{-A^2\tau f_s}{2}} + A\frac{T-\tau}{2T}e^{\frac{-A^2\tau f_s}{2}}\frac{-f_s}{\pi\tau^2}\frac{1}{2\sqrt{\frac{f_s}{\tau\pi}}} + A\frac{T-\tau}{2T}\sqrt{\frac{f_s}{\tau\pi}}\frac{-A^2 f_s}{2}e^{\frac{-A^2\tau f_s}{2}}$$
$$= \frac{-A}{2T}\sqrt{\frac{f_s}{\tau\pi}}e^{\frac{-A^2\tau f_s}{2}}\left(1+\frac{T-\tau}{2\tau}+\frac{T-\tau}{2}A^2 f_s\right) \quad (16)$$

Thus, the second derivative of U can be written as follows:

$$U"(\tau_i,\tau_j) = K\left(G'(\tau_i)-\frac{-A}{2T}\sqrt{\frac{f_s}{\tau\pi}}e^{\frac{-A^2\tau_i f_s}{2}}\right)$$
$$= \frac{-KA}{2T}\sqrt{\frac{f_s}{\tau\pi}}e^{\frac{-A^2\tau_i f_s}{2}}\left(2+\frac{T-\tau_i}{2}\left(A^2 f_s+\frac{1}{\tau_i}\right)\right) \quad (17)$$

It's obvious that K is a positive term, A is a positive if $P_{fi}(\epsilon,\tau_i)\leq 0.5$, and the other terms of (17) are all positives, thus the second derivative is negative and the function is quasiconcave.

B. Uniqueness of the Nash equilibrium

We turn now to examine the uniqueness of Nash equilibrium of the sensing-time game.

Proposition 2: The sensing-time game G admits a unique Nash equilibrium.

Proof: The game has a unique equilibrium if it satisfies the Rosen's conditions cited in [12].
If a concave game satisfies the dominance solvability conditions [12] given by:

$$-\frac{\partial^2 U_i}{\partial^2\tau_i}-\sum_{j\neq i}|\frac{\partial^2 U_i}{\partial\tau_i\partial\tau_j}|\geq 0 \quad (18)$$

Then it also satisfies Rosen's conditions.
For simplicity of the analysis We derive the utility function for 2 users, and check the sign of (23) using Maple software and found that is a positive term, thus the utility function satisfies the Dominance Solvability conditions and admits one Nash equilibrium. Simulation results eventually come to prove this result.

$$\frac{\partial(U_i(\tau_i,\tau_j))}{\partial\tau_i} = -\frac{\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_i f_s}\right)\right)P_i\left(1-P_j\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{T}+\frac{1}{2}\frac{(T-\tau_i)\,\mathrm{e}^{-A^2\tau_i f_s}Af_sP_i\left(1-P_j\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{\sqrt{\pi}\sqrt{\tau_i f_s}T} \quad (19)$$

$$\frac{\partial^2(U_i)}{\partial^2\tau_i} = -\frac{\mathrm{e}^{-A^2\tau_i f_s}Af_sP_i\left(1-P_j\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{\sqrt{\pi}\sqrt{\tau_i f_s}T}-\frac{1}{2}\frac{(T-\tau_i)\,A^3{f_s}^2\mathrm{e}^{-A^2\tau_i f_s}P_i\left(1-P_j\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{\sqrt{\pi}\sqrt{\tau_i f_s}T}-\frac{1}{4}\frac{(T-\tau_i)\,\mathrm{e}^{-A^2\tau_i f_s}Af_s{}^2P_i\left(1-P_j\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{\sqrt{\pi}\,(\tau_i f_s)^{\frac{3}{2}}T} \quad (20)$$

$$\frac{\partial(U_i(\tau_i,\tau_j))}{\partial\tau_j} = -\frac{1}{2}\frac{(T-\tau_i)\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_i f_s}\right)\right)P_iP_j\mathrm{e}^{-A^2\tau_j f_s}Af_s}{\sqrt{\pi}\sqrt{\tau_i f_s}T} \quad (21)$$

$$\frac{\partial^2(U_i)}{\partial\tau_i\partial\tau_j} = \frac{1}{2}\frac{\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_i f_s}\right)\right)P_iP_j\mathrm{e}^{-A^2\tau_j f_s}Af_s}{\sqrt{\pi}\sqrt{\tau_j f_s}T}-\frac{1}{4}\frac{(T-\tau_i)\,\mathrm{e}^{-A^2\tau_i f_s}A^2{f_s}^2P_iP_j\mathrm{e}^{-A^2\tau_j f_s}}{\pi\sqrt{\tau_i f_s}\sqrt{\tau_j f_s}T} \quad (22)$$

from (22) and (20) we can write:

$$-\frac{\partial^2 U_i}{\partial^2\tau_i}-\sum_{\mathrm{J}\neq i}|\frac{\partial^2 U_i}{\partial\tau_i\partial\tau_j}| = \frac{\mathrm{e}^{-A^2\tau_i f_s}Af_sP_i\left(1-P_i\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{\sqrt{\pi}\sqrt{\tau_i f_s}T}+\frac{1}{2}\frac{(T-\tau_i)\,A^3{f_s}^2\mathrm{e}^{-A^2\tau_i f_s}P_i\left(1-P_j\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{\sqrt{\pi}\sqrt{\tau_i f_s}T}+\frac{1}{4}\frac{(T-\tau_i)\,\mathrm{e}^{-A^2\tau_i f_s}Af_s{}^2P_i\left(1-P_j\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_j f_s}\right)\right)\right)}{\sqrt{\pi}\,(\tau_i f_s)^{\frac{3}{2}}T}+\Big|-\frac{1}{2}\frac{\left(1-\frac{1}{2}\,erfc\left(A\sqrt{\tau_i f_s}\right)\right)P_iP_j\mathrm{e}^{-A^2\tau_j f_s}Af_s}{\sqrt{\pi}\sqrt{\tau_j f_s}T}+\frac{1}{4}\frac{(T-\tau_i)\,\mathrm{e}^{-A^2\tau_i f_s}A^2{f_s}^2P_iP_j\mathrm{e}^{-A^2\tau_j f_s}}{\pi\sqrt{\tau_i f_s}\sqrt{\tau_j f_s}T}\Big| \quad (23)$$

IV. Distributed learning algorithm

In this section, we introduce an adaptive scheme to learn equilibria in the proposed Game. Learning is important in cognitive medium access control because the above payoff can be noisy due measurement, feedback errors or acknowledgment

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

errors. Learning will help us to understand the behavior of the users during the interactions. In the proposed game the sensing time represent a continuous action, a convenient class of learning scheme of our game is the class of COmbined fully DIstributed Payoff and Strategy learning (CODIPAS) for continuous actions. In the CODIPAS scheme each player seeks to learn her payoff functions as well as the associated equilibrium strategies. CODIPAS scheme, is well-adapted to cognitive radio networks for multiple reasons:

- CODIPAS is designed for random environment.
- It is only based numerical measurements and include noisy observation, measurement error and outdated measurements.
- The player does not need to know the others actions or others payoffs.

Let $\tau_{i,k}$ be the sensing time for user i at time k. $\tau^*_{i,k}$ is an intermediary variable. a_i, Ω_i and ϕ_i represent the amplitude, frequency and phase of the sinus perturbation signal respectively[17]. $U_{i,k+1}$represents the reward or the utility of user i at time k+1.

At each time instant k, each transmitter updates its sensing time $\tau_{i,k}$, by adding the sinus perturbation to the intermediary variable $\tau^*_{i,k}$ using equation (8),and makes the transmission using τ_i, k. Then, each transmitter gets a realization of the reward $U_{i,k+1}$ from its corresponding receiver at time k + 1 which is used to compute $\tau^*_{i,k+1}$ using equation (24).

The sensing time $\tau_{i,k+1}$ is then updated using equation (25). This procedure is repeated for the whole transmission window T. The algorithm is in discrete time and is given by:

$$\tau^*_{i,k+1} = \tau^*_{i,k} + \lambda_k l_i a_i sin(\Omega_i k^* + \phi_i) U_{i,k} \quad (24)$$

$$\tau_{i,k+1} = \tau^*_{i,k+1} + a_i sin(\Omega_i k^* + \phi_i) \quad (25)$$

Where $k^* := \sum_{k'=1}^{k} \lambda_k . \forall i$ should be selected to be small. For example $\lambda_k = \frac{1}{k+1}$. $\phi_i \in [0, 2\pi] \forall i, k \in Z$

Algorithm 1 CODIPAS for Continuous Sensing Time

1. Initialization :
Each secondary user i, initializes $\tau^*_{i,0}$ and transmits.
2. Learning pattern :
Each secondary user.
Observes the realization $U_{i,k}$.
Estimates $\tau^*_{i,k+1}$ according to Equation (24)
Updates $\tau_{i,k+1}$ using Equation (25).

V. Comparison with centralized optimal sensing time

In this section we propose a comparison of our distributed sensing method, with a centralized one. In centralized sensing, a central unit collects sensing information from SUs, identifies the available spectrum and broadcasts this information to other SUs or directly controls the CR traffic. The sensing results are combined in a central node. In the case of distributed sensing, cognitive nodes share information among each other but they make their own decisions when they have to determine which part of the spectrum they can use.

We can write the utility function for a centralized game with N SU as follow:

$$U(\tau) = \frac{T-\tau}{\tau} \left(1 - P_{fi}(\epsilon, \tau)\right) P_i \prod_{\jmath \neq i} \left(1 - P_j \left(1 - P_{fj}(\epsilon, \tau)\right)\right) \quad (26)$$

In this case the game can be expressed as:

$$\text{Find:} \qquad \tau = argmax(U(\tau)) \quad (27)$$

VI. Simulation results

In order to validate the theoretical ideas discussed above, simulation results are presented in this section,the primary signal is considered to be i.i.d complex shift keying modulated with bandwidth 6Mhz.

The parameters of simulations are set to be : T=100ms, $f_s = 6Mhz$, $P_{fi}(\epsilon, \tau_i) \leq 0.5$, $\phi_i = 0$, $a_i = 0.9$. The numerical setting could be tuned in order to make the convergence slower or faster. Each simulation setup is rum several times in order to smooth up the results and to eliminate the random behaviour effect during learning phase.

Fig. 2 shows that the SUs converge to an equilibrium sensing time, the sensing time is not the same for all users because SU does not have the same probability of access to the channel. As we can see in the Fig. 2 shows that the SUs converge to an equilibrium sensing time, the sensing time is not the same for all users because SU does not have the same probability of access to the channel. The simulations are repeated several times and the convergence is always to a unique point for a given set of parameters which assist our finding that the equilibrium is unique. A surprising feature is the observed correlation between transmit probability and sensing time. Explicitly, we notice that the lowest is the probability of access, the lowest is the sensing time !! This could be explained as follows. In order to respect licensed user (primary user) in terms of false alarm and detection probability, secondary users fine-tune their couple (transmit probability, sensing time). As a SU transmit with low probability, then even if it senses the channel for a brief period, this would not have a penalizing impact on PU activity. However, as a tagged SU becomes more aggressive (transmit at high probability), it needs to sense the channel during a longer period, so as it can catch the right PU's activity.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

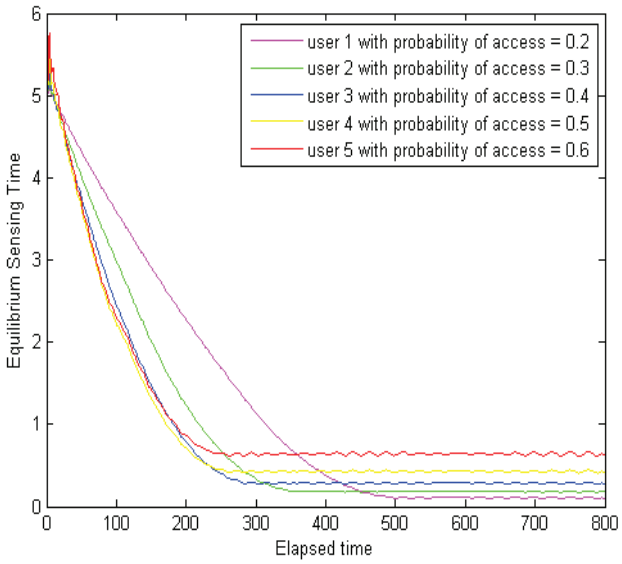


Fig. 2. Convergence and sensing time evolution for 5 users under $\lambda_k = \left(\frac{1}{k+1}\right)^{0.2}$.

Fig. 3 illustrates the convergence when the starting points are changed, we started here from different smaller points compared to Fig. 2, we notice that the users converge quickly to their equilibrium sensing time depending on their starting points, the more the starting points are small, the more the users converge quickly to their equilibrium sensing time. The users converge to the same equilibrium in Fig. 2 and Fig. 3 even if the starting points are changed.

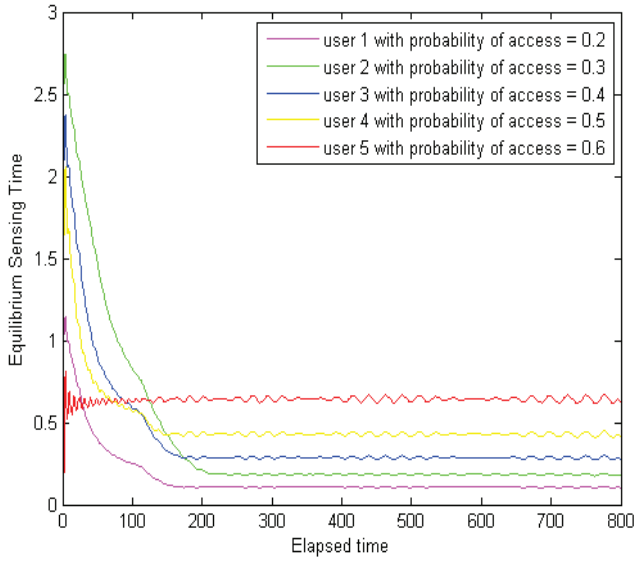


Fig. 3. Impact of the starting point on the convergence speed for 5 users under $\lambda_k = \left(\frac{1}{k+1}\right)^{0.2}$.

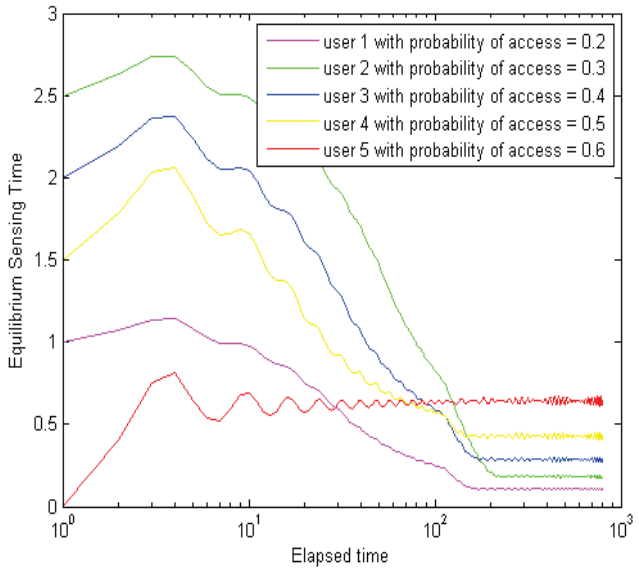


Fig. 4. Convergence in a logarithmic scale for 5 users under $\lambda_k = \left(\frac{1}{k+1}\right)^{0.2}$.

Fig. 4 is a logarithmic representation for Fig. 3 to highlight the start of the convergence that we can not see clearly in Fig. 3. The users are starting from different points compared to Fig. 2 and they converge to the same equilibrium sensing times.The uniqueness of the equilibrium sensing time is well illustrated in the case of user 5, we have seen in Fig. 2 that the user 5 converge to the higher equilibrium sensing time because of its higher probability of access, in Fig. 4 we can see that even if it starts from 0 ms it converges to the higher sensing time, which comes to prove the uniqueness of the equilibrium.

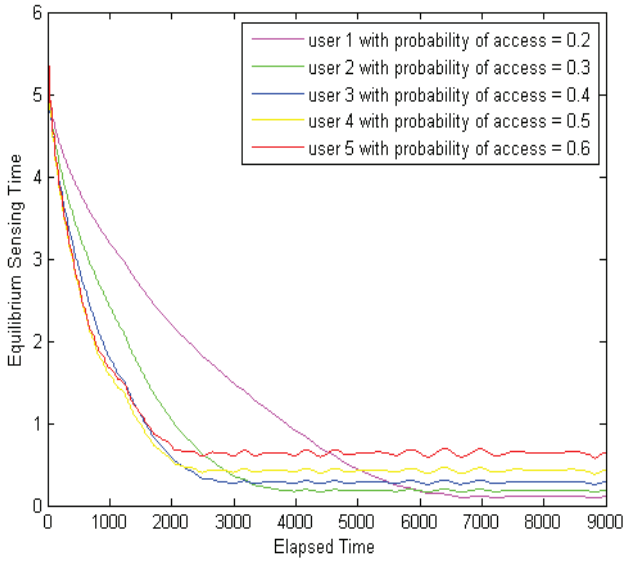


Fig. 5. Convergence and sensing time evolution for 5 users under $\lambda_k = \left(\frac{1}{k+1}\right)^{0.5}$.

Next, we depict in Fig. 5 the learning evolution for a speed

of convergence $\lambda\left(\frac{1}{k+1}\right)^{0.5}$. By comparing Fig. 5 with Fig. 2, where λ is set to a lower value $\left(\frac{1}{k+1}\right)^{0.2}$, we can see that the convergence is faster as λ increases. As expected the more the step is small, the more the convergence is slow, but the less fluctuations are strong. Thus, we need to define a trade-off between accuracy of convergence and speed of convergence.

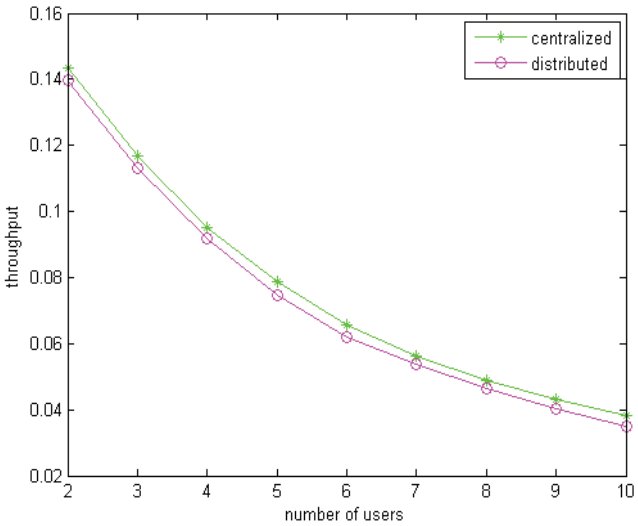


Fig. 6. comparison of the secondary users througput in both centralized and distributed method

In Fig. 6 we compare the throughput of secondary users in the proposed distributed methode, to a centralized one, as expected the centralized throughput is better, but as we can see in the figure the difference is minimal. The distributed method is more advantageous in the sense that there is no need for a backbone infrastructure, but as can be seen in Fig. 6 we lost some of the users throughputs.

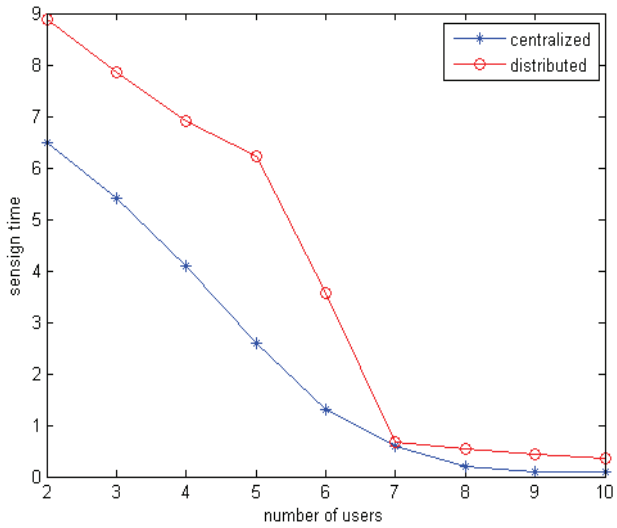


Fig. 7. equilibrium sensing time for distributed spectrum sensing vs centralized spectrum sensing

Fig. 7 shows a comparison between the equilibrium sensing time and the centralized one, the difference is slight especialy when the number of users increases.

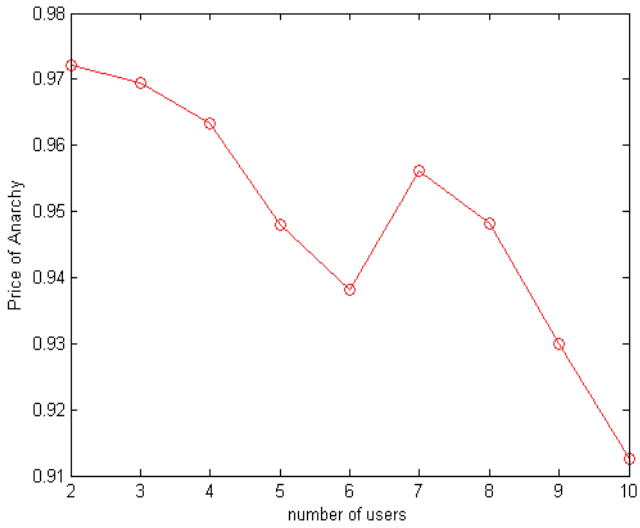


Fig. 8. Price of Anarchy

Fig. 8 represents the price of anarchy (POA) that measures the efficiency of a system, we measure the POA in this figure by computing the ratio of distributed throughput to the centralized one. the figure shows that the efficiency of the system degrades when the number of users increases, which can be explained by the selfish behavior of the users.

VII. CONSLUSION

In this paper, we proposed a distributed method for designing the sensing time in cognitive radio networks. We analyzed the sensing interaction among secondary users as a non-cooperative game. Then, we studied the equilibrium of this sensing time game and proved that there exist a unique equilibrium for the proposed game. Furthermore, we proposed a learning algorithm that is fully distributed and allows the cognitive users to learn jointly their payoffs and sensing time without use of any centralized controller. A comparison between the proposed solution and the centralized one. The distributed method is more advantageous in the sense that there is no need for a backbone infrastructure, the obtained results shows that the centralized method gives higher performances in terms of sensing time and throughput, thus a trade-off between the two solution must be defined.

REFERENCES

[1] S.Haykin. "Cognitive radio: brain-empowered wireless communications, IEEE J. Sel. Areas in Commune. vol. 23, pp. 201220, Feb 2004.
[2] J. Mitola, "Cognitive radio: An integrated agent architecture for software defined radio, Doctor of Technology, Royal Inst. Technol. (KTH), Stockholm, Sweden,2000.
[3] K.J.R. Liu and B. Wang, "Cognitive Radio Networking and Security: A Game Theoretical View ", Cambridge University Press,2010.
[4] D. Niyato, E.Hossain, and Z. Han, "Dynamics of Multiple-Seller and Multiple Buyer Spectrum Trading in Cognitive Radio Networks: A Game Theoretic Modeling Approach", IEEE Transactions on Mobile Computing, volume 8, no. 8, p.p. 10091022 ,August 2009
[5] J.Park, M.van der Schaar, "Cognitive MAC Protocols Using Memory for Distributed Spectrum Sharing Under Limited Spectrum Sensing ", IEEE Transactions on Communications 59(9): 2627-2637,2011
[6] H. Urkowitz, "Energy detection of unknown deterministic signals", Proceeding of the IEEE, Vol. 55, No. 4, pp. 523531 , April 1967
[7] Dandawat, A. V. and G. B. Giannakis, "Statistical tests for presence of cyclostationarity ", IEEE Trans. Signal Processing, 42(9), pp. 23552369.
[8] L. Lai, H. El Gamal, H. Jiang and H. Vincent Poor, "Cognitive Medium Access: Exploration, Exploitation and Competition ", IEEE Transactions on Mobile Computing, vol. 10, no. 2, pp. 239253, Feb. 2011.
[9] A. Anandkumar, N. Michael, A.K. Tang, and A. Swami, "Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret", IEEE JSAC on Advances in Cognitive Radio Networking and Communications, vol. 29, no. 4, pp. 781745, April 2011.
[10] M. Maskery, V. Krishnamurthy, and Q. Zhao, "Game Theoretic Learning and Pricing for Dynamic Spectrum Access in Cognitive Radio", in Cognitive Wireless Comm. Networks. Springer, 2007.
[11] H. Tembine, E. Altman, R. ElAzouzi, Y. Hayel "Evolutionary Games in Wireless Networks", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 40, issue 3, pp. 634646, 2010.
[12] P.Shi "Summary of What's Known about Convergence in Concave Games", April 2009.
[13] Y. C. Liang, Y. H. Zeng, E. Peh, and A. T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks", IEEE Transactions on Wireless Communications, vol. 7, no. 4, pp. 13261337 , April 2008
[14] A. Ghasemi and E. Sousa, "Optimization of spectrum sensing for opportunistic spectrum access in cognitive radio networks", in Proc. of 4th IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, USA , Jan 2007.
[15] R. W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria", International Journal of Game Theory 2, pp. 65-67. 1973
[16] D. Fudenberg and J. Tirole, "Game Theory", Cambridge, MIT Press.1991.
[17] A.Farhan Hanif, H.Tembine, M.Assaad, D.Zeghlache, "DISTRIBUTED STOCHASTIC LEARNING FOR CONTINUOUS POWER CONTROL IN WIRELESS NETWORKS", This paper appears in: Signal Processing Advances in Wireless Communications (SPAWC), 2012 IEEE 13th International Workshop, June 2012.
[18] M. Cardenas-Juarez and M. Ghogho, "Spectrum Sensing and Throughput Trade-off in Cognitive Radio under Outage Constraints over Nakagami Fading", IEEE Communication letters, 15(10), 2011.
[19] E. Sabir and T. Hamidou and M. Haddad, "Joint Strategic Spectrum Sensing and Opportunistic Access for Cognitive Radio Networks", IEEE GLOBECOM, December 2012.

VIII. AUTHORS PROFILE

Sofia BOUFERDA received the "Maitrise Sciences et Techniques" in "Informatique, Electronique, Electrotechnique et Automatique" from FSTM (Facult des sciences et techniques de Mohammedia) in 2005,she received the master degree in signal processing, telecommunications, image, networks and multimedia form Paris 13 university in 2006. In December 2010 she joined RITM laboratory at university Hassan II/ENSEM as a PHD student.

Essaid SABIR received the B.Sc. degree in Informatique Electronique Eletrotechnique et Automatique (2004) and the Diplome d'études Supérieures Approfondies (2007) from the Mohammed V University and National Institute of Post and Telecommunications (Rabat, Morocco). In 2010, he received the Ph.D degree in Networking and Computer Sciences from University of Avignon (France) and the Mohammed V University (Rabat, Morocco). He was contractual associate professor at University of Avignon, from 2009 to 2012.
He acts as a reviewer for professional publications, prestigious international journals and international conferences. He has/is been involved in several national and international/European projects. Currently, he is a full-time associate professor at National Higher School of Electricity and mechanics (ENSEM). His current research interests include performance evaluation, protocols design, ad hoc networking, flexible radio, stochastic learning, networking games, pricing and network neutrality.

Dr. Aawatif HAYAR received the "Agregation Genie Electrique" from Ecole Normale Suprieure de Cachan in 1992. She received the "Diplme d'Etudes Approfondies" in Signal processing Image and Communications and the degree of Engineer in Communications Systems and Networks from ENSEEIHT de Toulouse in 1997. She received with honors the Ph.D. degree in Signal Processing and Communications from Institut National Polytechnique de Toulouse in 2001. She was research and teaching associate at EURECOMs Mobile Communication Department from 2001 to 2010. Aawatif Hayar is currently with GREENTIC R&D Organization (Morocco) as General Secretary and with AH-Consulting (France) as expert in ICT field. She has also joined in 2011 the engineering school ENSEM at the University Hassan Casablanca in Morocco as Associate Professor. Aawatif Hayar was a Guest Editor of Elsevier Phycom Journal Special issue on Cognitive Radio Algorithms and System Design in 2009 and General Co-chair of Crowncom2010 and IW2GN2011. Aawatif Hayar Received best student paper award at CogArt2010 and has a patent on "Process for sensing vacant bands over the spectrum bandwidth and apparatus for performing the same based on sub space and distributions analysis".

Mounir RIFI was born in Fez, Morocco in 1962. He is with EST (Ecole Superieure de Technologie) at the University Hassan II of Casablanca as Professor of Higher Education, since September 2000 and as Assistant Professor EST University of Fez, October 1987 2000. He is also member of Doctoral Studies Centre "Engineering Sciences", Head of Research Team "Networks & Telecoms" and Director of the Research Laboratory: RITM (Networks, Computer, Telecom and Multimedia). Prof. Rifi obtained his PhD in Electronics, May 1987 (University of Lille - France). He is Board member of GREENTIC Casablanca association, founder of the Mediterranean Association of Telecommunications and publisher and founder of the Mediterranean Telecommunications Journal. His research activities covers several area of research fields such as electromagnetic waves propagation , ElectroMagnetic Compatibility, RFID, Transmission Lines, Smart antenna, Sensor Networks, Computer Networks.

A Neural Approach for Analysis of Lymphocytes in Detection of Rheumatoid Arthritis using Center Approximation Method

Murugan .V
Department of Computer science and Engineering,
Sri Venkateswara College of Engineering,
Chennai, India

Thivakaran .T.K
Department of Computer science and Engineering,
Sri Venkateswara College of Engineering,
Chennai, India.

***Abstract*— Rheumatoid Arthritis is an autoimmune disease that mainly affects joints in the human body. It is a chronic disease that causes stiffness, pain, swelling and limited motion of various joints which leads to the erosion of bones at the joints. As treatments are currently available only to delay or stop the erosion and not to regain the eroded joints, it is important to detect its occurrence at the onset of aberration. Early detection of Rheumatoid Arthritis involves analysis of Lymphocytes present in human blood. Manual analysis of Lymphocytes is a long term process which needs an expert hematologist for continuous microscopic assessment of blood smear which is costly and time consuming. Digital image processing plays a vital role in the field of medical diagnosis. The proposed work aims to minimize the expenditure of the assessment of Lymphocytes with the help of digital image processing techniques. 105 samples of blood smear images containing Lymphocytes are collected from various laboratories out of which 60 images are affected and 45 images are not affected by rheumatoid arthritis. In order to reduce the noise from these images, initially they are enhanced using Weiner Filter. To locate the region of interest the enhanced images are subjected to thresholding. After thresholding image is allowed to segmentation using bounding box technique in order to separate the nucleus of the Lymphocyte from the smear image. The segmented nucleus is used to extract the required features: area, perimeter and circularity. Here the circularity value is obtained using center approximation method. These features are given to a single layer perceptron which achieved a classification accuracy of about 97.50%.**

Keywords- Adaptive thresholding, Circularity of Lymphocytes.

I. INTRODUCTION

Rheumatoid arthritis (RA) is one of the most common and serious forms of arthritis. RA is a chronic, autoimmune disease, mainly characterized by inflammation of the lining, or synovium, of the joints. Without early disease-modifying treatment, progressive and irreversible joint damage can occur resulting in lifelong functional impairment. Rheumatoid arthritis can have a severe impact on health, quality of life, socio-economic status, and is associated with premature death. Findings of population-based studies show RA affects 0·5–1·0% of adults in developed countries [1]. Estimates of the frequency of rheumatoid arthritis vary depending on the methods used to ascertain its presence. Incidence ranges from 5 to 50 per 100 000 adults in developed countries and increases with age. Early diagnosis is paramount as prompt initiation of disease-modifying treatment can limit the joint damage that occurs in rheumatoid arthritis [2]. However, early identification of rheumatoid arthritis is difficult because of the dearth of perceptible disease-specific onset features.

There is no specific sign, symptom or diagnostic feature that distinguishes early rheumatoid arthritis from other forms of inflammatory arthritis. Methods to detect rheumatoid arthritis include clinical assessment, imaging and laboratory tests [3]. Accurate diagnosis of RA has not been proved to be possible by in-situ laboratory testing. Attempts have been made to correlate data taken from various tests on an individual. These tests cannot be completely sidelined as futile; they can narrow down possibilities, suggest a prognosis, evaluate symptoms and monitor the side effects of drugs [4]. One such test is the shape feature analysis of lymphocytes. This requires analysis of blood smear images by an expert hematologist under laboratory conditions, which is an iterative and complex process.

Manual analysis of Lymphocytes is a time consuming process and it may also lead to human errors. This necessitates the need for a computerized tool to analyze the shape features of lymphocytes more precisely. Digital image processing plays a vital role in the development of computerized tools to analyze digital images. Advances in technology have made medical imaging a worthwhile tool in the diagnosis of various diseases and analysis of various parts of the body. The central idea of this work is to develop a tool that assists a hematologist in analyzing the shape features of lymphocytes in order to predict whether it is inflamed or not.

II. RELATED WORKS

Innumerable researches have been carried out in the field of medical image processing like analysis of blood smear images, radiographic images, MRI (Magnetic Resonant Imaging) images, Tomography images, Ultrasound images, Echocardiography images, Thermograph images, etc. A lot of intelligent techniques have been developed to diagnose

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

various diseases such as malaria, Sickle cell Anemia and Acute Myeloid Leukemia through the analysis of Red Blood Cells (RBCs) and White Blood Cells (WBCs) in a blood smear image. In this paper, we mainly concentrate on the analysis of WBCs which requires a good segmentation algorithm which then separates our Region of Interest (ROI) from the entire blood cell image. The accuracy of the segmentation algorithm has a great impact on the final results of the analysis. Various approaches are available for the segmentation of ROI from the entire image.

There are five different types of WBCs, namely: Lymphocyte, Monocyte, Eosinophil, Basophil and Neutrophil. In order to classify WBCs, the information about the nucleus is alone sufficient [5]. By applying the method specified by Nipon Theera-Umpon et.al [5], the problem of cells that touch each other can be eliminated. The segmentation of nucleus alone from the smear image is easier than segmenting the entire cell. This can be achieved by using mathematical morphology of WBCs. Procedure for automatic detection of malaria from a microscopic blood smear image is presented by J.Somasekar, et.al [6]. A two stage colour segmentation strategy with fuzzy clustering is used for separating WBCs from other blood components [7]. After extracting the Leukocytes, Lymphocytes are identified from the segmented image as a subclass. Various features such as fractal features, shape features and other texture features are extracted from the sub class of Lymphocytes for detection of Acute Leukemia.

SP. Chokkalingam et.al [8] analyzes the irregularities of Lymphocytes in a blood smear image using area, perimeter, integrated density, solidity and circularity. It is an arduous procedure involving standard mathematical calculations and approximations. To overcome the complexity, a new approach for shape feature calculation involving centroid approximation and distance from centre of the nucleus to the periphery has been used.

III. METHODOLOGY

Rheumatoid Arthritis is an auto immune disease which affects the Lymphocytes present in human blood. It is observed that the Lymphocytes of persons who got affected by RA are deformed from its original shape [8]. The proposed work is to analyze the irregularities in shape of Lymphocytes present in human blood smear image using simple image processing techniques. The complete work flow diagram of the proposed method is shown in figure 1. Form the figure it is clear that the given test image is preprocessed using Wiener filter, then nucleus of the Lymphocyte is segmented using adaptive thresholding, Canny edge detector and bounding box. Features are extracted from the segmented nucleus and given to a single layer perceptron, a simple artificial neural network which matches the extracted features with the previously created dataset of known sample images and gives the output as infected or non infected sample. Above mentioned steps are briefly discussed in the following sub section.

The content in this Section is organized as follows: Part A describes about the acquisition of blood smear image, part B briefs about the preprocessing operations applied to the acquired blood smear image, part C explains about the Segmentation of Lymphocytes from the smear image, part D briefs the process of extracting the required features from the segmented Lymphocytes and Finally part E explains about the classification process using single Layer perceptron.

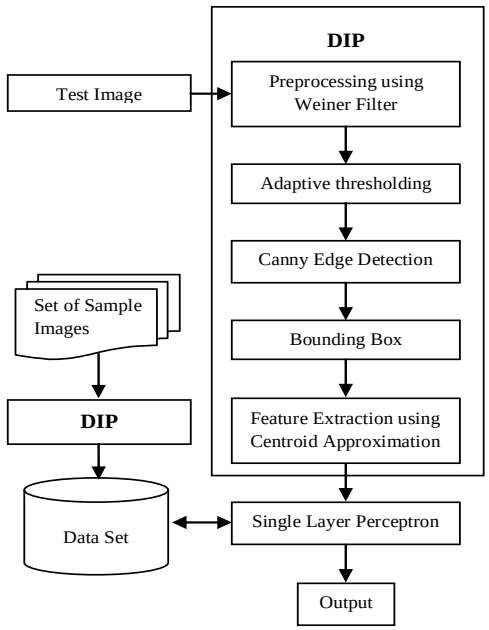


Figure 1: Work flow Diagram

A. *Image Acquisition:*

Images of Leishmann stained blood smears were collected Bharat Scans Laboratory, Chennai. 100x oil immersion views of these images are obtained using Samsung Carl Zeiss Microscope. Captured images were 640 pixels x 480 pixels bitmap images.

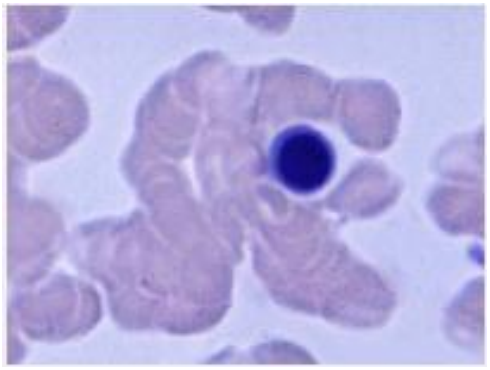


Figure 2: Sample blood smear image that contain a normal Lymphocyte

Captured images were 640 pixels x 480 pixels bitmap images. A dataset is formed using 105 such images for training the single layer perceptron. All images taken must have only Lymphocytes in it; Images with other white blood

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

cells must be ignored. Figure 2 shows a sample blood smear image with RBCs and Lymphocytes.

B. Preprocessing using Weiner Filter:

The aim of preprocessing is to remove unwanted noise from the obtained image and to transform the image in such a manner so that it is easy to do further processing. Any microscopic image is highly subjected to Gaussian noise. To choose the appropriate filter, an analysis of various filters has been made on images subjected to Gaussian noise. From the analysis report shown in Table I, it is clear that a Weiner filter shows good performance over Gaussian noise. As processing a color image is more complex, initially all obtained images are converted into grey scale images and then the Weiner filter is applied to it before entering into the next stage of processing. A Weiner filter is a linear filter which optimizes the given image in terms of low mean square error. It involves linear estimation of desired signal another related sequence. When image shown in figure 2 is subjected to preprocessing using a Weiner filter the resulting image will be as show in figure 3.

TABLE I. ANALYSIS OF FILTERS OVER GAUSSIAN NOISE

Filter/ Parameters	Mean Square Error	PSNR	Normalized Cross Correlation	Normalized Absolute Error
Mean	98.5693	28.1934	0.9938	0.0441
Median	115.4250	27.5078	0.9990	0.0523
Weiner	39.2802	32.1891	0.9956	0.0275
Gaussian	76.1603	29.3135	0.9936	0.0384

C. *Image Segmentation:*

It is the process of separating the Region of Interest (ROI) from the given image. In this work, nucleus of the Lymphocytes is our ROI. There are a lot of methods which allows segmentation of ROI from the large image which includes thresholding, Edge detection, region growing methods, split-and-merge method and histogram based methods. In this paper, a combination of Adaptive thresholding, Canny Edge detection and bounding box technique is used to separate the ROI. The purpose of thresholding is to remove lighter part of the image and to show only the darker region. Canny edge detector is used to deduce the boundary of the nucleus and Bounding box is use to separate the ROI as sub image from the original image.

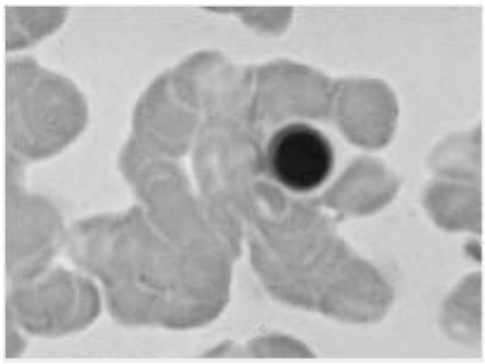


Figure 3: Preprocessed blood smear image

Initially the obtained peripheral blood smear image contains RBCs and Lymphocytes. Due to Leishmann staining RBCs become lighter and the Nucleus of Lymphocytes become darker. For any given grey scale image, thresholding helps to obtain a binary image based on the threshold value given. A local threshold method is adapted based on the nature of the image.

The preprocessed image is a grey scale image in which each pixel value range from 0 to 255, where a 0 represents black and 255 represents white pixel. Thresholding is a method when applied to grey scale image, all pixels which have a value greater than the given threshold value are represented as black pixel i.e. 0 and pixels which are lower than the threshold value is represented a white pixel i.e. 1.The proposed method is to increase the threshold value until the lighter part of the preprocessed image becomes white and the darker region i.e. the nucleus of a Lymphocyte becomes black. So if we apply adaptive thresholding to the image shown in Figure 3, the resulting image will contain only the nucleus which is darker than other regions in the image and is shown in Figure 4. As a result of thresholding the blood smear image we can eliminate the unwanted RBC's and cytoplasm of the Lymphocytes from it.

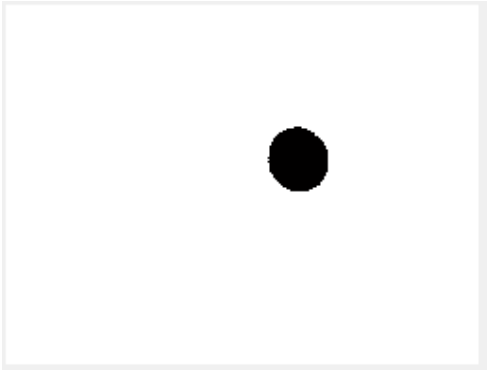


Figure 4: Thresholded Blood smear image having only the Nucleus of the Lymphocyte

In the process of feature extraction, in order analyze the shape properties of the nucleus of Lymphocytes: perimeter and circularity, we need the edge of the nucleus. Edge Detection is a mechanism that allows us to reduce the amount of data in an image, preserving the structural properties in an image. There are many edge detection algorithms like Canny, Canny-Deriche, Differential, Sobel, Prewitt and Roberts. Out of these the canny edge detector [9] is the simplest edge detection algorithm and hence it is used to get the edge or boundary of the nucleus in a thresholded image. It uses a multi-stage algorithm to detect the edge boundary of the nucleus accurately. The stages of a canny edge detector are as follows: Smoothing the given image in order to remove any noise in the image, Finding the gradients such that the edges are marked, Non-maximum suppression is done to mark local maxima as edges, Double thresholding to determine the potential edges and final edges are determined by suppressing the edges that are not connected to strong edges. When canny

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

edge Detection is applied to the image shown in figure 4, the resulting image will be as shown in figure 5.

Figure 5: Edge of the Nucleus given by canny edge detector

To separate the nucleus alone from the remaining image a bounding box technique is used. Bounding box algorithm is used to separate various connected components in an image. It locates the connected component using the maximum and minimum (x, y) co-ordinates in an image and allows us to crop the required portion out of the image. Here in Figure 5, the only connected component is the nucleus of the Lymphocyte.

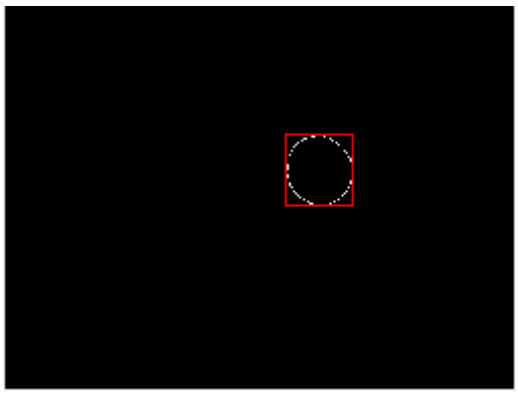


Figure 6: Nucleus of Lymphocyte Bounded using Bounding Box technique

Hence if we apply bounding box algorithm to it will result in an image shown in figure 6. We can apply crop operation to the image in Figure 6, and separate the nucleus alone from its background. Thus the ROI i.e. the Nucleus of the Lymphocyte is obtained as a sub image which is used for further processing.

D. Feature Extraction:

Feature Extraction is the process reducing the dimensionality of input. All images contain large amount to information in it, which is more redundant. To reduce the number of inputs to a classifier an image is converted into a set of feature vectors using feature extraction process. Lymphocytes have specific features based on the size and circularity it is decided whether the sample is infected by rheumatoid arthritis or not. Inthis paper area and perimeter mentions size features of Lymphocytes and Circularity of a lymphocyte is approximated based on the distribution of number of lines on a specified region. These features are calculated using the following method. Area is calculated from the thresholded image, i.e. number of dark pixels in figure 4 gives the area of the nucleus in number of pixels. The next feature perimeter is calculated from the edge detected image. In figure 5, the boundary of the nucleus is shown in white pixel and hence to calculate the perimeter we can count the number of white pixel in that image.

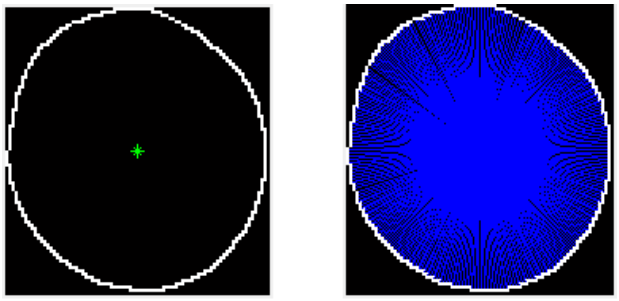


Figure 7: (a) Centroid of Nucleus; (b) Lines Drawn from centroid to edge in order to calculate the distance from center of the nucleus to the edge

Circularity is a measure that how much circular the given object i.e. the nucleus is. To predict the circularity a new method involving centroid approximation is used. The algorithm for this feature vector extraction is as follows:

Step 1: Locate the centroid of the nucleus in the cropped image as shown in Figure 7(a).

Step 2: Calculate the minimum and maximum length from centroid to the edge pixel.

Step 3: Divide the interval into ten equal intervals and initialize a counter for each region.

Step 4: Now draw lines from centroid to all edge pixels as shown in Figure 7(b) in order to calculate the distance from centroid to all edge pixel one by one and increment the counter value of the respective region.

Table II and III shows the values of above mentioned features for ten samples of normal Lymphocytes and abnormal lymphocytes respectively.

E. Classification:

Image classification is the most important and final step in most image analysis applications. Image classification can be either supervised or unsupervised. A single layer Perceptron is the simplest binary classifier which maps the given input x to either 0 or 1. In this paper, our work is to classify whether the given sample is inflamed or not. As the classification problem is Linear, a single layer perceptron is chosen as the classifier. It is implemented with Hard Limit as transfer function and LEARNP as learning function. Cyclical Weight/Bias rule is chosen as the training rule for the neural network.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE II. FEATURES FOR NORMAL LYMPHOCYTE IMAGE

Area	Perimeter	Number of Lines in Region									
		1	2	3	4	5	6	7	8	9	10
4520	290	24	30	25	35	38	27	29	46	28	8
5260	307	39	28	55	53	29	16	32	16	16	23
5722	338	14	23	16	15	22	25	68	99	38	18
4640	300	8	9	21	55	58	40	41	39	20	9
4720	305	20	28	15	17	35	30	27	42	59	32
4670	302	27	28	27	21	36	59	43	30	14	17
4859	308	23	15	30	48	36	37	38	22	38	21
5023	314	23	33	23	41	39	52	34	30	20	19
4795	305	35	31	21	15	20	56	48	26	31	22
5116	323	56	23	18	31	33	50	30	25	22	35

TABLE III. FEATURES FOR ABNORMAL LYMPHOCYTE IMAGE

Area	Perimeter	Number of Lines in Region									
		1	2	3	4	5	6	7	8	9	10
7716	372	23	35	35	15	29	78	66	38	31	22
6793	353	17	24	21	28	48	62	61	42	30	20
6721	369	37	21	22	22	30	67	42	29	35	64
4956	298	13	9	13	13	21	23	25	51	55	75
8525	408	45	33	58	80	23	75	17	26	24	27
11466	579	75	63	49	119	47	42	41	43	34	66
9370	456	26	9	10	10	9	33	127	40	85	103
5998	339	36	23	16	16	13	42	67	47	35	44
7291	377	43	20	23	29	90	80	14	18	31	29
5909	339	26	18	14	13	12	13	36	66	97	44

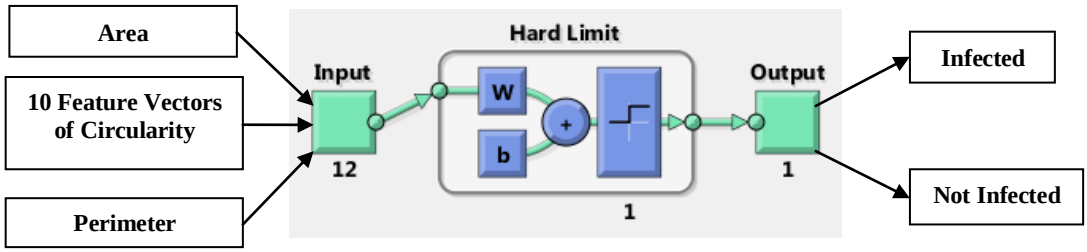


Figure 8: Structure of Perceptron

105 samples of peripheral blood smear images out of which 60 infected and 45 non infected samples were taken and features of each image is extracted using the above mentioned method. Each image is reduced into twelve feature vectors and these feature vectors are given to the neural network for training.

IV. RESULTS AND CONCLUSION

The trained network is tested with 40 samples of peripheral blood smear images of patients who are infected and not infected with rheumatoid arthritis. The network shows good performance detecting 39 samples accurately. The accuracy of the classifier is measured in terms of recall and Precision. Recall marks the ability of the algorithm to detect whether the sample is infected or not and Precision marks the success of the classifier at excluding non-infected Lymphocytes. These values are expressed in terms of False Positive (FP), False Negative (FN) and True Positive (TP) is as follows:

$$recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

```
Correctly Classified Instances          39               97.5    %
Incorrectly Classified Instances         1                2.5    %
Total Number of Instances               40

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  Class
                 0.967     0          1         0.967   inflamated
                 1         0.033      0.909     1       noninflamated
Weighted Avg.    0.975     0.008      0.977     0.975
```

Figure 9: Accuracy of the classifier

The accuracy of the classifier is depicted in figure 9, form the results it is clear that the classifier is robust and more accurate. With this tool we can analyze the probability of a patient being affected by rheumatoid arthritis or not. The future enhancement can be made in the segmentation algorithm, in order to accurately segment Lymphocytes from samples of peripheral blood smear images with all other types of white blood cells

REFERENCES

[1] David L Scott, Fredrick Wolfe, Tom W. J. Huizinga, "Rheumatoid Arthritis". Lancent. Vol. 376, pp. 1094-1108, 2010.

[2] Kim J. M., Weisman M. H. , "When does rheumatoid arthritis begin and why do we need to know?" Arthritis Rheum, Vol. 43(3), pp. 473–484, March 2000.

[3] D. Symmons, G.Turner, R.webb, P.Asten, E.barrett, M.Lunt, D.scott and A.silman, "The prevalence of Rheumatoid arthritis in United Kingdom: new estimates for a new century", Rheumatology , Vol. 41, pp. 793-800, 2002.

[4] M. E. a. McNeil, The First Year--Rheumatoid Arthritis: An Essential Guide for the Newly Diagnosed, in Da Capo press, 2005.

[5] Nipon Theera-Umpon and Sompong Dhompongsa, "Morphological Granulometric Features of Nucleus in Automatic Bone Marrow White Blood Cell Classification", IEEE Transactions on Information Technology in Biomedicine, Vol. 11(3), pp. 353-359, May 2007.

[6] J.Somasekar, B.Eswara Reddy, E.Keshava Reddy and Ching-Hao Lai, "An Image Processing Approach for Accurate Determination of Parasitemia in Peripheral Blood Smear Images", IJCA Special Issue on Novel Aspects of Digital Imaging Applications (DIA), Vol. 1 pp. 23–28, 2011.

[7] Subrajeet Mohapatra, Dipti Patra and Sanghamitra Satpathi, "Image Analysis of Blood Microscopic Images for Acute Leukemia Detection", International Conference on Industrial Electronics, Control and Robotics, IECR, pp. 215-219, Dec 2010.

[8] SP. Chokkalingam, K.Komathy, "Analysis of Irregular shapes of Lymphocytes in automatic detection of rheumatoid arthritis", Europen journal of Scientific research, Vol. 82(1), pp. 46-57. 2012.

[9] John Canny, "A computational approach to edge detection", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-8(6), pp. 679–698, Nov. 1986.

[10] Tycko D. H., S. Anbalagan, H. C. Liu, and L. Ornstein, "Automatic leukocyte classification using cytochemically stained smears," J Histochem Cytochem, Vol. 24, pp. 178-194, Jan 1976.

[11] Petya em. Pavlova, Krassimir P. Cyrrilov and Ivan N. Moumdjiev, "Application of HSV colour system in the identification by colour of biological objects on the basis of microscopic images", Computerized Medical Imaging and Graphics, Vol. 20(5), pp. 357-364, Sep-Oct 1996.

[12] D. Anoraganinrum "Cell Segmentation with Median Filter and Mathematical Morphology Operation", International Conference on Image Analysis and Processing, pp. 1043-1046, Sep 1999.

[13] Bharanidharan. T, and D. K. Ghosh, "A Two Dimensional Image classification Neural Network for Medical Images", European Journal of Scientific Research,Vol.74(2) , pp. 286-291, 2012.

[14] Yampri .P, C. Pintavirooj, S. Daochai and S. Teartulakarn, "White Blood Cell Classification based on the Combination of Eigen Cell and Parametric Feature Detection", IEEE Conference on Industrial Electronics and Applications, pp. 1-4, May 2006.

[15] Vishnu V. Makkapati and Raghuveer M. Rao, "Segmentation of Malaria Parasites in Peripheral Blood Smear Images", IEEE International conference on Acoustics, Speech and Signal Processing, pp. 1361-1364, April 2009.

[16] Di Ruberto, Andrew Dempster, Shahid Khan and Bill Jarra, "Analysis of Infected Blood Cells Images Using Morphological Operators", Journal Of Image And Vision Computing, Elsevier, vol. 20(2), pp. 133-146, Feb 2002.

[17] Rafael C. Gonzalez, "Digital Image processing", Preason Education India, Edition 3, 2009.

[18] Bacusmber, James .W, Gose, Earl .E, "Leukocyte Pattern Recognition", IEEE Transactions on Systems, Man and Cybermetics, Vol. SMC-2(4), pp. 513-526, Sep1972.

[19] Anggraini .D, Nugroho .A .S, Pratama .C, Rozi .I. .E, Iskandar .A .A, Hartono .R .N, " Automated Status Identification of microscopic images obtained from malaria thin Blood Smears", International Conference on Electrical Engineering and Informatics, pp. 1-6, July 2011.

The Use of Cuckoo Search in Estimating the Parameters of Software Reliability Growth Models

Dr. Najla Akram AL-Saati
Software Engineering Dept
College of Computer Sciences & Mathematics
Mosul, Iraq

Marwa Abd-AlKareem
Software Engineering Dept
College of Computer Sciences & Mathematics
Mosul, Iraq

***Abstract*— this work aims to investigate the reliability of software products as an important attribute of computer programs; it helps to decide the degree of trustworthiness a program has in accomplishing its specific functions. This is done using the Software Reliability Growth Models (SRGMs) through the estimation of their parameters. The parameters are estimated in this work based on the available failure data and with the search techniques of Swarm Intelligence, namely, the Cuckoo Search (CS) due to its efficiency, effectiveness and robustness. A number of SRGMs is studied, and the results are compared to Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO) and extended ACO. Results show that CS outperformed both PSO and ACO in finding better parameters tested using identical datasets. It was sometimes outperformed by the extended ACO. Also in this work, the percentages of training data to testing data are investigated to show their impact on the results.**

Keywords- Software Reliability; Growth Models; Parameter estimation; Swarm Intelligence; Cuckoo Search

I. INTRODUCTION

Software reliability has gained a huge importance recently due to the fact that it takes a long time for a company to build up a reputation for reliability, but only a short time to be acknowledged as "unreliable" subsequent to shipping a faulty product. Repeated appraisal of new product reliability and the constant reliability control of every shipped product are critical requirements in today's competitive business arena [1].

It has been found throughout the continual practice in this area that the process of proving or testing cannot assure absolute dependability neither on the product nor in its correctness. Therefore a metric is required to act as a measure for determining the degree of program correctness. One of the most widely used quality metrics is software reliability. It is usually measured using analytical models whose parameters are estimated from real failure data. [2]

Software Reliability can be defined as [3]:

"The probability that a system or product will perform in a satisfactory manner for a given period of time when used under specified operating conditions in a given environment."

To measure reliability in an effective manner, many issues have to be carefully considered, such as the accurateness of time to failure, the time to failures sequence, and failure mode data [4]. Evaluating software reliability requires many techniques, and nearly all of these techniques depend upon constructing prediction models with the ability to predict upcoming faults under diverse testing situations [5]. These models are generally named Software Reliability Growth Models (SRGMs).

In the past four decades, software reliability growth models began to be introduced for estimating the reliability of software programs. The nature of most of these models were non linear, which made the estimation of the parameters hard to accomplish using basic methods.[6] this idea opened the door for other methods to take part in estimating the parameters for non linear models, and the interest in Evolutionary Computation began to have effect in solving different Software Engineering problems.

Reliability growth models were deeply studies throughout the literature and their construction was investigated by many authors[2][7]. Numerous models were taken into consideration, and of the intensively used in the literature are those with two parameters, such as: Logarithmic [8], Exponential (Goel-Okumoto Model) [9], Power [10], S-Shaped (Yamada S-Shaped Model) [11][12] and Inverse Polynomial models [13].

Research in Software Engineering has recently been witnessing an enormous advance with the use of Evolutionary Computational methods. This is specially observed in finding acceptable solutions to prediction, estimation and optimization problems in Software Engineering. [14]

In this work, a number of models are introduced throughout the investigation. A detailed study of reliability and its growth models is introduced along with the problem of estimating their parameters, and then the methodology of the cuckoo search algorithm is presented to signify its role in estimating the parameters of the growth models. The results are compared with those obtained by [14] using the Exponential (Goel-Okumoto), the S-shaped and Power models whose parameters were estimated using PSO. In addition the results are also to be compared with those achieved by ACO and extended ACO in [6] using Exponential (Goel-Okumoto), S-shaped, Power, and M-O models. Comparisons indicate the efficiency of the CS algorithm.

II. RELATED WORK

Parameter estimation in software reliability models were traditionally solved using the Maximum Likelihood method or the Least Square method; unfortunately these two methods are

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

not suitable for non linear software reliability growth models. Therefore different solution methods and algorithms such as GAs, GP, NN, Fuzzy Logic, and Swarm Intelligence, have all had their share at trying to solve various problems in the Software Engineering area.

SRGMs were frequently studied and investigated throughout the literature; and here are some of these studies:

Kuo et al. in (2001) offered a framework for modeling software reliability by the use of various testing-efforts and fault detection rates. [7]. In (2002), Okamura et al. estimated a mixed Software Reliability Models using the Expectation-Maximization (EM) algorithm. [15] In addition, Okamura, Murayama and Dohi also used the (EM) algorithm in (2004) and developed a unified parameter estimation method for discrete software reliability models.[16]

By (2005), Huang [17], made a performance analysis of SRGMs with testing effort and change-point. After that, Ando, Okamura and Dohi introduced another work in (2006) about estimating Markov modulated software reliability models by the use of EM Algorithm [18]. In the same year, Sheta used PSO to solve the parameter estimation problem for the exponential, power and S-Shaped models [14]. During (2008), Bokhari, Quadri, and Khan, proposed the Exponetiated Weibull SRGM with various testing-efforts and optimal release policy with a performance analysis.[19] In (2009), Ohishi, Okamura, and Dohi presented a Non-Homogenous Poisson Process (NHPP) to develop an estimation algorithm of Gompertz software reliability model [20]. Throughout (2011), Quadri, Ahmad, and Farooq proposed a scheme for constructing a SRGM based on NHPP with generalized exponential testing – effort and optimal software release policy [21]. In the same year, Miglani and Rana proposed a greedy approach for ranking of different software reliability growth models [22].

Lately, in (2012), Shanmugam and Florence made a comparison of parameter best estimation methods and showed that ACO was the best among them [23]. At the same year, they made an improvement on ACO and compared it to their previous work [6].

Recently in (2013), a computational methodology based on weighted criteria was presented to the problem of performance analysis of various NHPP models [24]. In the same year, Okamura et.al, proposed a model based on a mixed gamma distribution, the estimation method was based on Bayesian estimation and the estimation algorithm was described by Markov chain Monte Carlo method with grouped data [25].

III. SOFTWARE RELIABILITY GROWTH MODELS (SRGMS)

A. *Definitions and Classification:*

Throughout the premature stages of developing and prototyping complex systems, reliability did not commonly meet customer requirements [1]. Reliability Models have usually been used to accurately evaluate and predict the behavior and performance of software reliability. In the 1970's, studies in software reliability models became more attracting and achieved greater advances, many reliability models have already been put into use.

Analysis methods of software reliability can either be white box or black box reliability analysis; these two differ from each other in that white box reliability analysis take into account the internal structure of the software to estimate its reliability, whereas black box (software reliability growth models) regards the software as a monolithic undividable unit by using failure data that took place in the middle of external interactions.[26] Fig. (1) shows the difference between white and black box reliability analysis and their relationship with software development stages.

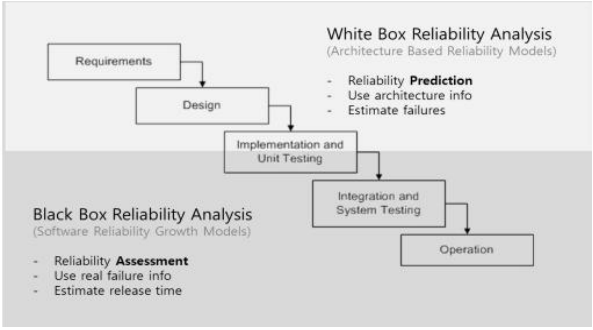


Figure 1. Black Box and White Box Reliability Analysis [26]

Software reliability models can generally be considered as static and the dynamic models. Static Models apply the modeling and analysis of program logic on the same code, while Dynamic models observe the temporary behavior of debugging process throughout the phase of testing [23].

There are many opinions related to the classification of reliability models. Basically there are two types [27]:

- **Defect Density Models** that try to predict reliability from design parameters. They employ code characteristics like code lines, loop nesting, external references, input/outputs, and others to estimate the number of defects in software.
- **Software Reliability Growth Models** which use test data to predict software reliability. These models intend to statistically correlate defect detection data with known functions like, for example, an exponential function. When that correlation is acceptable, the known function can be used to predict future behavior.

The use of reliability growth models (the focus of this work) for predicting software reliability signifies a huge challenge for software testing. predicting the number of faults inhabited in software programs gives a considerable assists in denoting the day for software release and control project resources (people and money) [28]. Most software reliability growth models provide work for the estimation of two or three parameters, these models include the predictable number of failures in the software, and the initial failure intensity.

B. *Characteristics and Notations:*

The characteristics of Software Reliability Growth Model's should be satisfied for the Software Reliability Model [29]. These Characteristics are as follows:

- SRGM can be viewed as a product of a cumulative density function and a positive constant.

$$H(t)=a\left[1-\exp\left(-\int_0^1 d(x)dx\right)\right]=aG(t), \quad (1)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Where:
H (t) is mean value function.
G (t) is Cumulative Density Function.

- The fault detection rate must be finite, and G(t) must meet the condition that the corresponding failure rate function is finite.

$$h(t) = d(t)[a - H(t)] = = ag(t), \quad (2)$$

Where
g(t) =dG (t)/dt, the probability density function associated with G (t).

$$d(t) = ag(t)\frac{ag(t)}{a - aG(t)} = \frac{g(t)}{1 - G(t)} = r(t) \quad (3)$$

Where:
r(t) the failure rate function associated with G (t).
h(t): is intensity function.

- It is necessary to appropriately represent the right tail of g(t) behavior to achieve a good reliability prediction. The right tail of SRGM associated with g(t) should be heavy.

The terms in Table (I) are to be used in defining the Models in the next subsections.

TABLE I. TERMS USED FOR DEFINING SRGMS

Term	Definition
μ(t)	Denotes the mean failure function, i.e., the expected number of failures observed over a period of time t.
λ(t)	Denotes the failure intensity function, i.e., failure rate
a	The initial estimate of the total failure recovered at the end of the testing process.
b	Represents the ratio between the initial failure intensity λ0 and total failure.
NHPP	The Non Homogenous Poisson Process: provides probability that the number of failures at a time t will have a particular value.

C. Models Employed in this Work:

In this work a number of models are considered, in particular those that are frequently and commonly referenced in the literature.

The models studied in this work are:

1) Exponential Model (Goel-Okumoto G-O) [9]: Goel-Okumoto model is recognized as a finite failure model that can be modeled as NHPP. The Prediction of the model can be given as:

$$\mu(t) = a(1 - e^{-bt}) \quad (4)$$

$$\lambda(t) = abe^{-bt} \quad (5)$$

The number of faults to be detected (a) is handled as a random variable for which the observed value depend on the test and other environmental factors. This is fundamentally different from other models that treat the number of faults as a fixed unknown constant [30].

2) The Power Model [10]:

This model has the objective of computing the reliability of hardware systems during testing process. It is also based on the NHPP [31]. The equations (6) and (7) rule the relationship between the time t and both μ (t) and λ (t).

$$\mu(t) = at^b \quad (6)$$

$$\lambda(t) = abte^{b-1} \quad (7)$$

3) The Yamada Delayed S-Shaped Model [11][12]:

The Delayed S-Shaped Model is of the gamma distribution class. But the number of failures per time period is a Poisson type with the use of the classification scheme of Musa and Okumoto rather than considered as Binomial. This model describes the software reliability process as a delayed S-shaped model. The model represents a learning process since some improvement was added to the exponential model based the growing experience of the project team. This model is also a finite failure model [14]. The system equation for μ (t) and λ (t) are:

$$\mu(t) = a(1 - (1 + bt)e^{-bt} \quad (8)$$

$$\lambda(t) = ab^2t^{-bt} \quad (9)$$

4) Musa-Okumoto Logarithmic Model [8]:

This is continuous time-independently distributed inter failure time model. This is a modification of the J-M model where a geometrically decreasing hazard function is introduced, considering that only one fatal error is removed during each debugging interval. Faults are not removed until the occurrence of a fatal one at which time the accumulated group of faults is removed. The hazard function after a restart is a fraction of the rate which was attained when the system crashed [23]. The Prediction Model form is given as:

$$\mu(t) = a \ln(1 + bt) \quad (10)$$

D. Parameter Estimation:

Estimating the parameters problem for nonlinear systems can be stated and formulated as a function optimization problem. The purpose is to discover a set of parameters that provide the best fit to a measured data based on a specific type of function to be optimized. Such parameters are found using a search procedure in the space of values specified in advance. Searching techniques are bound to the complexity of the search space, and the use of Gradient search might find local minimum solution but not optimal ones. Stochastic search algorithms, on the other hand, such as Evolutionary Algorithms present a more reliable functionality in estimating models' parameters. [14]

Swarm Intelligence has been successfully used to provide efficient mechanisms in searching for solutions to this problem, of these mechanisms, PSO and ACO were used as stated previously and in this work Cuckoo Search (CS) is applied to search for better estimates associated with the parameter values of SRGMs and its efficiency and performance is compared to those obtained by PSO and ACO.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV. CUCKOO SEARCH (CS):

New meta-heuristic search algorithms are rapidly increasing under the paradigm of swarm intelligence, resembling the intelligence inhabited in creatures from nature. Algorithms such as Particle Swarm Optimization (PSO), Ant colony optimization (ACO), Bee Colony Optimization (BCO) Fish Schools, and many others have become well known in solving various problems. Of these algorithms, Cuckoo Search (CS) [32] which was developed by Yang and Deb in 2009, has proven to be very promising through preliminary studies and conformed giving more robust and precise results than PSO and ABC (Artificial Bee Colony).[33]

A. *Cuckoo's Inspiring Behavior:*

Some species of Cuckoo are interesting, such as the Ani and Guira, as they have a strange habit of laying their eggs in public nests; they may also remove others' eggs to boost the hatching probability of their own eggs [34]. Quite a number of species engage the obligate brood parasitism by laying their eggs in the nests of other host birds (often other species). There are basically three types of brood parasitism:

- Intra-specific brood parasitism.
- Cooperative breeding.
- Nest takeover.

Some host birds can go on a direct conflict with the intruding cuckoos. If a host bird finds that the eggs are not its own, it will either throw away these alien eggs or simply dump its nest and build a new one in another place. Some cuckoo species such as the new world brood-parasitic Tapera have evolved somehow that female parasitic cuckoos are often very expert in the mimicry in the egg's color and pattern of a few chosen host species. This decreases their eggs' probability of being abandoned and therefore enhances their reproduction activity. [32]

Also taking into account that parasitic cuckoos habitually choose a nest where the host bird just laid its own eggs. Usually, the cuckoo eggs hatch a little earlier than their host eggs. When the first cuckoo baby bird is hatched, its first instinct act will be to throw out host eggs, this will increase its share of food that is supplied by the nest's host bird. Various Studies in cuckoo behavior also indicate that a cuckoo baby bird can mimic the call of host baby bird to obtain more feeding opportunity.[32]

In this paper, CS will be applied to parameter estimation of SRGMs to conduct a more detailed study of its characteristics and to verify it against benchmark datasets. After that a discussion is establish the unique features of Cuckoo Search and propose topics for further studies.

B. *Lévy Flights:*

Animals habitually search for food in a random or quasi-random style. Generally, the foraging path of an animal is in fact a random walk; this is because the next move depends on the current location or state and the transition probability to the next location. The chosen direction depends implicitly on a probability which can be mathematically modeled. For instance, various studies on animals and insects have shown that their flight behavior demonstrates the typical characteristics of Lévy flights [35].

Reynolds and Frye [35] explored fruit flies (Drosophila Melanogaster), and indicated that they explore their landscape using a series of straight flight paths punctuated by a sudden 90o turn, leading to a Lévy-flight-style intermittent scale-free search pattern. Afterwards, Studies on human behavior were conducted as well and also showed the typical feature of Lévy flights. Many other things can also be related to Lévy flights, such as light [36]. That is why such behavior has been applied to optimization and optimal search, where preliminary results show its promising capability [37].

C. *Search Strategy for Cuckoos:*

For the ease of description, the following three idealized rules are used [37]:

- In each time, every cuckoo lays one egg in a randomly chosen nest.
- Only best nests having high quality eggs (solutions) will continue to the next generations;
- The available host nests are fixed in number. A host can discover an alien egg with a probability pa $\in$ [0, 1]. When discovered, the host bird can either throw the egg away or dump the nest to build a totally new one in another location.

```
Cuckoo Search via Lévy Flights
Begin
    Objective function f(x), x = (x1, ..., xd) T
    Generate initial population of n host nests xi (i = 1, 2, ..., n)
    While (t <MaxGeneration) or (stop criterion)
      Get a cuckoo randomly by Lévy flights
        Evaluate its quality/fitness Fi
      Choose a nest among n (say, j) randomly
      If (Fi > Fj),
            Replace j by the new solution;
      End
      A fraction (pa) of worse nests is abandoned and new ones are built;
      Keep the best solutions (or nests with quality solutions);
      Rank the solutions and find the current best
      End while
      Post-process results and visualization
      End
```

Figure 2. Pseudo code of the Cuckoo Search (CS) [32]

Based on these three rules, the basic steps of the Cuckoo Search (CS) can be presented as the pseudo code shown in Fig. 2 [32]. To give more simplicity, the last assumption can be approximated by a fraction pa of the n nests being replaced by new nests (with new random solutions at new locations). The fitness can be defined in the same way as done in Genetic Algorithms. When a new solution $x_i^{(t+1)}$ is generated for the i'th cuckoo, a Lévy flight is done as the following:

$$x_i^{(t+1)} = x_i^{(t)} + \propto \oplus \text{ Lévy}(\lambda), \qquad (11)$$

where:

$\alpha > 0$ is the step size (it should be related to the scales of the problem at hand). Most of the time, it is used as $\alpha = O(1)$.

$\oplus$ is the entry-wise multiplications.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Lévy flight is used to conduct a random walk drawn from a Lévy distribution for large steps:

$$\text{Lévy} \sim u = t^{-\lambda}, (1 < \lambda \leq 3), \quad (12)$$

This has an infinite variance with an infinite mean. The successive jumps/steps of a cuckoo basically form a random walk which obeys a power-law step-length distribution with a heavy tail.

V. Parameter Estimation Based On CS:

Based on the characteristics of the SRGMs, the cuckoo search is implemented as follows:

- To solve the problem of parameter estimation, the principal of Root Mean Square Error (RMSE) is used as in (13):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}[m(t) - \mu(t)]^2}, \quad (13)$$

  The total actual discovered failure number in time t, is m(t), the predicted failure number by SRGM are expected failure number by time t is μ(t) .

- each nest carries a pair of eggs representing solution parameters (a, b), at each time a cuckoo is chosen according to levy flight, its fitness is evaluated and compared to the fitness of a randomly chosen nest, the best fitness is kept.
- A fraction (pa) of nests is abandoned and new nests are built.

VI. Tests and Results

To test the efficiency of the search algorithm employed in this work, two types of testing are conducted. In the beginning, comparisons are made with previous results gained using PSO (with three models) and ACO (with four models) using the same related datasets. Then, the splitting of Datasets between training and test sets is investigated to show the impact of their sizes on results; this is done using three models and one benchmark dataset. Table (II) shows the parameter settings for the cuckoo search employed in this work for all the experiments.

TABLE II. Parameter settings for the Cuckoo Search Algorithm

Parameter	Value
Lower and Upper bounds for Parameter a	[0.00001 - 2000]
Lower and Upper bounds for Parameter b	[0.00001 – 1]
Number of Cuckoos	1
Number of Nests	10
Number of Eggs	2
Number of iterations (Generations)	100
Alpha	0.01
Discovery rate	0.25

A. Experimental Data used in this work

Datasets used in this work are chosen in accordance to those referenced by other researchers with which the comparisons were made; the first group of datasets (compared with PSO) is taken from [14] for Data1, Data2, and Data3. The second group (compared with ACO and extended ACO) is selected from The Software Reliability Dataset which was compiled by John Musa of Bell Telephone Laboratories [38] for Project 2, Project 3, and Project 4.

B. Comparison with other Swarm Algorithms:

The cuckoo search algorithm's methodology is compared to that of PSO using the same datasets employed in his work, and Table (III) signifies the training and testing percentages as divided by Sheta [14] for exponential (EXP), power (POW), and Delayed S-Shaped Yamada Model (DSS), the results show the clear improvement achieved in testing the parameters for the specified models for Data1. Tables (IV) and (V) shows the comparisons made using Data2 and Data3for the same models.

These results were accomplished using 100 iterations when compared to that used by [14], where 1000 iteration were required to achieve the given PSO results.

TABLE III. Comparison with PSO using Dataset (Data1)

Search / Model	*RMSE - Training 70%*		*RMSE - Testing 30%*	
	PSO	***CS***	***PSO***	***CS***
EXP(G-O)	20.2565	34.0933	119.4374	16.8945
POW	22.2166	44.8663	152.9372	33.6623
DSS	15.9237	32.6376	26.3015	10.9945

TABLE IV. Comparison with PSO using Dataset (Data2)

Search / Model	*RMSE - Training 70%*		*RMSE - Testing 30%*	
	PSO	***CS***	***PSO***	***CS***
EXP(G-O)	24.9899	33.2311	80.8963	14.2998
POW	32.3550	47.0571	149.9684	56.6807
DSS	20.8325	27.9159	17.0638	11.8833

TABLE V. Comparison with PSO using Dataset (Data3)

Search / Model	*RMSE - Training 70%*		*RMSE - Testing 30%*	
	PSO	***CS***	***PSO***	***CS***
EXP(G-O)	12.8925	13.5404	13.6094	8.9523
POW	11.9446	13.0886	14.0524	13.4669
DSS	18.5807	13.6634	47.4036	15.1916

The methodology is further compared to ACO [23] and extended ACO [6] using the same datasets used in their work, they used the whole sets of data for training. Tables (VI), (VII), and (VIII) indicate the training percentages for Goel-Okumoto (G-O), power (POW), Delayed S-Shaped Yamada Model (DSS), and the Musa-Okumoto (M-O) model for Projects 2, 3, and 4.

Results in Table (VI) show that using Project2 with (G-O, POW, and DSS) models, CS outperformed ACO but not EX-ACO. But for the (M-O) model, CS gave the worst results.

TABLE VI. Comparison with ACO and Extended ACO (Project2)

Search / Model	**Project2 – Training 100%**		
	ACO	***Ex-ACO***	***CS***
G-O	60.0371	**28.5891**	41.7971
POW	52.8854	**34.0521**	45.9783
DSS	52.8854	**33.0461**	42.2256
M-O	26.0385	**17.359**	41.7732

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

For Project3, CS outperformed both ACO and EX-ACO for all models as illustrated in Table (VII). Using Project4, Table (VIII) indicates that CS surpassed both ACO and EX-ACO for (G-O, POW, and DSS) models. As for the M-O model, CS performed better than ACO but not better than EX-ACO. Fig. 3 to 5 depicts the data in Tables (VI, VII, and VIII) to indicate the differences among the three datasets used for the same models.

TABLE VII. COMPARISON WITH ACO AND EXTENDED ACO (PROJECT3)

Search / Model	Project3 – Training 100%		
	ACO	*Ex-ACO*	*CS*
G-O	71.5489	34.0709	**21.7256**
POW	57.5801	47.5814	**15.5885**
DSS	57.5801	48.4914	**22.4944**
M-O	36.1891	24.126	**19.5448**

TABLE VIII. COMPARISON WITH ACO AND EXTENDED ACO (PROJECT4)

Search / Model	Project4 – Training 100%		
	ACO	*Ex-ACO*	*CS*
G-O	71.4015	35.0007	**25.7682**
POW	53.2234	34.2645	**28.1951**
DSS	53.2234	35.2635	**25.7294**
M-O	33.1728	**22.1152**	26.4575

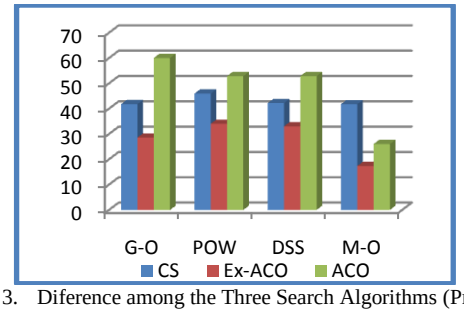


Figure 3. Diference among the Three Search Algorithms (Project2)

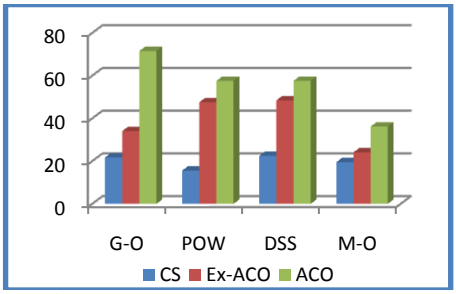


Figure 4. Diference among the Three Search Algorithms (Project3)

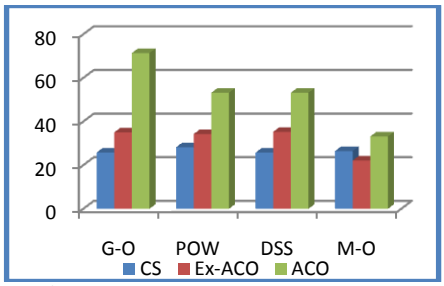


Figure 5. Diference among the Three Search Algorithms (Project4)

As a result, CS can perform better than PSO in all cases considered in this study. For ACO, it was noticed that CS has achieved better results in most cases, and for some cases it also surpassed EX-ACO. This is largely due to the nature of the datasets used in correlation with the model employed in the testing of parameters.

C. The Impact of Training and Testing Data:

Through the process of parameter estimation, the datasets used can either be used just for training, or it can be divided into two sets: Training and Testing. This division can have a large influence on results, and that is way it is investigated in this subsection.

The study included the three datasets (Data1, Data2, and Data3) used in the previous subsection. Tables (IX), (X), and (XI) show the impact of each model using the three datasets and five different percentages for training and testing for models (EXP, POW, and DSS). From the results, it can be seen that increasing the training data percentage over testing data, will enhance testing results and worsen training results, and vise versa. This is quite logical, as feeding a large amount of data to the training process forces the search procedure (cuckoo search here) to find very few but excellent solutions (proved by the corresponding testing process). When smaller amounts of data are used for training, they become unsatisfactory, and a large amount of testing data is not required.

To overcome this problem of overestimation and underestimation, a balanced point is chosen such as (70%, 30%) to give training reasonable amounts of data to achieve good training and adequate data to the testing process.

TABLE IX. IMPACT OF TRAINING AND TESTING PERCENTAGES (DATA1)

Model	EXP	POW	DSS
Training 90%	38.062	61.940	41.833
Testing 10%	8.811	6.481	2.747
Training 80%	36.37	54.77	35.77
Testing 20%	10.703	16.335	7.113
Training 70%	**34.093**	**44.866**	**32.638**
Testing 30%	**16.895**	**33.662**	**10.995**
Training 60%	32.877	39.858	30.985
Testing 40%	22.253	48.956	21.555
Training 50%	27.443	27.877	24.760
Testing 50%	31.539	73.668	36.336

TABLE X. IMPACT OF TRAINING AND TESTING PERCENTAGE (DATA2)

Model	EXP	POW	DSS
Training 90%	40.374	62.153	30.843
Testing 10%	3.490	7.574	6.209
Training 80%	36.841	56.576	28.637
Testing 20%	7.126	25.381	10.598
Training 70%	**33.231**	**47.057**	**27.916**
Testing 30%	**14.299**	**56.680**	**11.883**
Training 60%	29.876	39.468	25.648
Testing 40%	33.402	86.050	45.975
Training 50%	26.377	33.046	24.907
Testing 50%	68.858	125.906	71.220

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE XI. IMPACT OF TRAINING AND TESTING PERCENTAGE (DATA3)

Model	EXP	POW	DSS
Training 90%	18.539	19.525	17.229
Testing 10%	1.613	3.256	5.158
Training 80%	15.299	17.311	15.489
Testing 20%	3.916	8.287	10.588
Training 70%	**13.540**	**13.089**	**13.663**
Testing 30%	**8.952**	**13.467**	**15.191**
Training 60%	12.484	12.986	11.874
Testing 40%	16.476	16.578	33.636
Training 50%	11.823	12.132	10.673
Testing 50%	34.240	20.012	60.926

Also in this work, the number of iterations (generations), the number of nests (population size) and the probability (pa) were varied to find the best suitable settings. Number of iterations was varied from (50 to 1000) and the value 100 iteration was quite capable of achieving the best results. Nests of (10, 15, 20, 25, 50, 100, 150, and 500) were tried and the sufficient value was found to be (10). As for the probability pa, values of (0, 0.01, 0.02, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, and 0.7) were used and the best setting was found to be (0.25).

VII. CONCLUSIONS AND FURTHER RECOMMENDATIONS

It has been found through this work that the CS algorithm can be successfully applied to find good and acceptable solutions to the problem of parameter estimation of Software Reliability Growth Models. Models considered in this work are: the Exponential, Power, S-Shaped, and M-O models. The search strategy of the cuckoo can efficiently navigate throughout the search space of the problem and locate very good solutions using fewer iterations and smaller populations.

First, the Results were compared to PSO, ACO, and Extended ACO; the results clearly outperformed PSO, they were better than ACO in almost all cases, but were sometime worse than the extended ACO.

Second, the testing and training data sizes were investigated, pairs of (training, testing) were considered using three datasets they were: (90%, 10%), (80%, 20%), (70%, 30%), (60%, 40%), and (50%, 50%). It is concluded that increasing the percentage of training data, makes training hard and testing very simple and not sufficient enough. On the contrary, when percentage of training is decreased, the training data becomes insufficient to train and the testing data becomes very large and unnecessary. Thus a counterbalance point is chosen in the middle (70%, 30%) to give training a justified amount of data to accomplish good training and fair enough data to the testing process.

As for further recommendations, other swarm intelligent methods can be applied and compared; this may uncover more suitable parameter estimation methods. Future work might also include the application of other evolutionary search methods to construct new SRGMs that can assess the reliability more adequately.

REFERENCES

[1] C. Croarkin, P. Tobias. Nist/Sematech "E-Handbook of Statistical Methods", http://www.itl.nist.gov/div898/handbook, 2012.

[2] A. L. Goel. "Software Reliability Models: Assumptions, Limitations, and Applicability", IEEE Transactions on Software Engineering, Vol. 11, No. 12, pp: 1411–1423, 1985.

[3] B. S. Blanchard. "Logistics Engineering and Management". Englewood Cliffs: Prentice Hall, Inc, 1992.

[4] L. J. Hogge. "Effective Measurement Of Reliability Of Repairable Usaf Systems". Master of Science in Systems Engineering in the Air Force Institute of Technology. 136p, 2012.

[5] S. Yamada, S. Osaki. "Optimal Software Release Policies with Simultaneous Cost and Reliability Requirements", European J. Operational Research, pp: 46–51, 1987.

[6] L.Shanmugam, L. Florence. "An Improved ACO Technique for Parameter Estimation". In the European Journal of Scientific Research, ISSN 1450-216X Vol. 89 No 1, pp:101-108. © Euro Journals Publishing, Inc. 2012.

[7] S.Y. Kuo, C.Y. Hung, and M.R. Lyu. "Framework for modeling Software Reliability, Using Various Testing-Efforts and Fault Detection Rates", IEEE Trans. Reliability. Vol. 50, No.3, pp: 310-320, 2001.

[8] J. D. Musa, K. Okumoto, "A Logarithmic Poisson Execution Time Model for Software Reliability Measurement", Proc. 7th int'l Conf. Software Engineering. pp: 230-238, 1984.

[9] A.L. Goel, K. Okumoto, "Time-Dependent Error Detection Rate Model for Software Reliability and Other Performance Measures". In IEEE Transactions on Reliability, Vol.28, pp: 206-211, 1979.

[10] L. H. Crow, "Reliability for Complex Repairable Systems, Reliability and Biometry", SIAM, pp: 379–410, 1974.

[11] S. Yamada, M. Ohba, S. Osaki. "S-Shaped Reliability Growth Modeling for Software Error Detection". IEEE Trans. Reliability, 32(5), pp: 475-484, 1983.

[12] S. Yamada, M. Ohba, S. Osaki. "S-Shaped Software Reliability Growth Models and Their Applications," IEEE Trans. Reliability, pp: 289–292, 1984.

[13] B. Littlewood and J. L. Verall. "A Bayesian Reliability Model with A Stochastically Monotone Failure Rate", IEEE Trans. Reliability, Vol. 23, pp: 108–114, 1974.

[14] A. Sheta. "Reliability Growth Modeling for Software Fault Detection Using Particle Swarm Optimization", IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada. pp: 3071- 3078, 2006.

[15] H. Okamura, Y. Watanabe, T. Dohi. "Estimating Mixed Software Reliability Models Based on The EM Algorithm". Proc. Of the 2002 Int. Symposium Empirical Software Engineering. pp: 69-78, 2002.

[16] H. Okamura, A. Murayama, T. Dohi, "EM algorithm for Discrete Software Reliability Models: A Unified Parameter Estimation Method". Proceedings of the eighth IEEE International Symposium on High Assurance Systems Engineering. pp:219-228. 2004.

[17] C.Y. Huang. "Performance Analysis of Software Reliability Growth Models with Testing Effort and Change-Point", Journal of Systems and Software, Vol. 76, pp: 181-194, 2005.

[18] T. Ando, H. Okamura, T. Dohi, "Estimating Markov Modulated Software Reliability Models via EM Algorithm". In the Proceedings of the 2nd IEEE International Symposium on Dependable, Automatic and Secure Computing, 2006.

[19] N. Ahmad, M. U. Bokhari, S. M. K. Quadri, M. G. M. Khan. "The Exponetiated Weibull Software Reliability Growth Model with various testing-efforts and optimal release policy: a performance analysis". In International Journal of Quality and Reliability Management. Vol. 25 Iss: 2, pp: 211 – 235. 2008.

[20] K. Ohishi, H. Okamura, and T. Dohi, "Gompertz Software Reliability Model: Estimation Algorithm and Empirical Validation". In Journal of Systems and software, Volume 82, Issue 3. Pp: 535-543, 2009.

[21] S. M. K. Quadri, N. Ahmad, S.U. Farooq, "Software Reliability Growth Modeling with Generalized Exponential Testing – Effort and Optimal

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Software Release Policy". In the Global Journal of Computer Science and Technology. Volume 11. Issue 2. pp: 27-42. 2011

[22] N. Miglani, and P. Rana, "Ranking of Software Reliability Growth Models using Greedy Approach". In the Global Journal of Business Management and Information Technology. Volume 1, Number 2, pp. 119-124. © Research India Publications. 2011.

[23] L. Shanmugam, L.A. Florence, "Comparison of Parameter Best Estimation Method for Software Reliability Models". In the International Journal of Software Engineering & Applications (IJSEA), Vol.3, No.5. pp:91-102. 2012.

[24] M. Anjum, Md. Asraful Haque, A. Nesar."Analysis and Ranking of Software Reliability Models Based on Weighted Criteria Value". In I.J. Information Technology and Computer Science, 02, pp: 1-14, 2013.

[25] H. Okamura, T. Hirata, and T. Dohi. "NHPP-Based Software Reliability Model with Mixed Gamma Distribution". In Proceedings, The 2nd International Conference on Software Technology, SoftTech, ASTL Vol. 19, pp. 240 - 243, © SERSC 2013.

[26] D. Hong. "A White Box Reliability Prediction Technique for Supporting Tracability", Master Thesis. 40p, 2010.

[27] Z. Al-Rahamneh, M. Reyalat, A. F. Sheta, S. Bani-Ahmad, S. Al-Oqeili. "A New Software Reliability Growth Model: Genetic-Programming-Based Approach". In Journal of Software Engineering and Applications, 4, pp: 476-481. 2011

[28] K. Okumoto, A. L. Goel. "Optimal Release Time for Software System Based on Reliability and Cost Criteria", J. of Systems and Software, pp: 315–318, 1980.

[29] R. Jiang "Required Characteristics for Software Reliability Growth Models". In World Congress on Software Engineering IEEE 2009. DOI 10.1109/WCSE.2009.157

[30] H. Kan. Stephen. "Metrics and Models in Software Quality Engineering", Second Edition. Addison Wesley. ISBN: 0-201-72915-6. 560p, 2002.

[31] M. Krasich. "Power Law Model, Correct Application in Reliability Growth Do the Cumulative Times Indeed Always Add Up?". In Conference: Reliability and Maintainability Annual Symposium - RAMS, pp: 1-7, 2012.

[32] X.S. Yang, S. Deb, "Cuckoo search via Lévy flights". In: Proc. of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), December 2009, India. IEEE Publications, USA, pp: 210-214, 2009.

[33] P. Civicioglu , E. Besdok. "A Conceptual Comparison of the Cuckoo-search, Particle Swarm Optimization, Differential Evolution and Artificial Bee Colony Algorithms". In Artificial Intelligence Review. Volume 39, Issue 4, pp: 315-346. 2013.

[34] R. B. Payne, M.D. Sorenson, K. Klitz. "The Cuckoos", Oxford University Press. 2005.

[35] A. M. Reynolds, M. A. Frye. "Free-flight Odor Tracking in Drosophila is Consistent with An Optimal Intermittent Scale-Free Search", PLoS One, 2, e 354. 2007.

[36] P. Barthelemy, J. Bertolotti, D. S. Wiersma. "A Lévy Flight for Light", Nature, 453, 495-498. 2008.

[37] X.S. Yang, S. Deb, "Engineering Optimisation by Cuckoo Search". f. Int. J. Mathematical Modelling and Numerical Optimisation, Vol. 1, No.4, pp: 330-343, 2010.

[38] https://sw.thecsiac.com/databases/sled/swrel.php.

Toward a New Approach for Modeling Dependability of Data Warehouse System

Imane Hilal
RITM Lab., CED Engineering Sciences ENSEM
ESTC, Hassan II University
Casablanca, Morocco

Nadia Afifi
RITM Lab., Computer Engineering Department
ESTC, Hassan II University
Casablanca, Morocco

Reda Filali Hilali
RITM Lab., Computer Engineering Department
ESTC, Hassan II University
Casablanca, Morocco

Mohammed Ouzzif
RITM Lab., Computer Engineering Department
ESTC, Hassan II University
Casablanca, Morocco

***Abstract*—The sustainability of any Data Warehouse System (DWS) is closely correlated with user satisfaction. Therefore, analysts, designers and developers focused more on achieving all its functionality, without considering others kinds of requirement such as dependability's aspects. Moreover, these latter are often considered as properties of the system that will must be checked and corrected once the project is completed. The practice of "fix it later" can cause the obsolescence of the entire Data Warehouse System. Therefore, it requires the adoption of a methodology that will ensure the integration of aspects of dependability since the early stages of project DWS. In this paper, we first define the concepts related to dependability of DWS. Then we present our approach inspired from the MDA (Model Driven Architecture) approach to model dependability's aspects namely: availability, reliability, maintainability and security, taking into account their interaction.**

Keywords-component; Data Warehouse System; Model Driven Architecture ; Dependability; Availability; Reliability, Security, Maintainability.

I. INTRODUCTION

Data Warehouse Systems (DWS) are specially used by decision makers to analyze the status and the development of their organization *[1]*, based on a large amount of enterprise information integrated from heterogeneous Data Sources *[2]*. This information is organized following a multidimensional structure to make exploration and analysis easier for users. The DWS's architecture is composed from several layers: (i) Heterogeneous Data Sources (DS), (ii) ETL (Extraction/ Transformation/ Loading) process which extract and transform data from these DS, and load the information into DW ,(iii) Data Warehouse repository, (iv) Restitution Tools that analyze the data in OLAP way.

The final goals to implement the DWS are: (i) To evaluate complex queries without causing severe impact on the sources; (ii) To increase data availability and decrease response-time for OLAP queries; (iii) To provide correct historical trends for state-oriented data; (iv) To protect and secure the crucial business information; (v) To back up modifications and insure evolution and maintenance. Those goals involve the guarantee of non-functional requirements such as availability, reliability, security, confidentiality, integrity and maintainability which are encompassed in the dependability's attributes [3].

To meet those goals, we suggest an approach that spans over dependability's attributes. Our contribution presents, on the one hand, the advantage of considering these attributes from the early stages of the DW project. On the other hand, it provides models to the designers and developers in order to realize these functionalities in respect to the Model Driving Architecture's (MDA) principles.

The remainder of this paper is organized as follows: section 2 will present the related work on the development of dependable DWS; section 3 will give an overview of DW, dependability aspects and MDA; section 4 will introduce our approach through which we develop our models for DWS dependability. An example of implementation is shown in section 5. Finally, in section 6 we present our conclusion and future work.

II. RELATED WORK

In DW's literature, dependability's attributes have been neglected or they have been presented as a second class type of requirements, and have also been considered as the other non-functional requirements. But the experiences show that capturing non-functional requirements without mapping them into the conceptual model may provoke loss of information [4]. Only few works have specifically addressed this issue. In

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

particular, Pain &castro [5] proposed (DWARF) the Data Warehouse Requirements Definition approach that adapts a requirement engineering process for requirements definition and management of DW. Their approach demands particular attention to non-functional requirements that are captured through an extension of the NFR framework. The same authors have provided an exhaustive classification of non-functional requirements that must be addressed in the development of DWS [6].

A different proposal comes from V. Peralta, A.Illarze, R. Ruggia, [7] who have suggested modeling the non-functional requirements through guidelines that are not directly related with the conceptual model. Instead, non-functional requirements have been used during logical design, where most of the choices related to the performance and security have already been taken.

E. Soler, V. Stefanov, J. N. Mazón, J. Trujillo, E. Fernández-Medina, M. Piattini, [8] proposed a security requirement model for DW based on the MDA approach[9]. This approach was extended by Carlos Blanco [10] who has developed a secure DW, including security issues in each stage of the development process. This is the only approach that considers security as a non-functional requirement from the early stages of the DW development.

On the other hand, some aspects of dependability such as reliability, availability, security and maintainability have been widely discussed in the DWS[11]. However, their interaction and integration were rarely discussed [11, 12].

III. DEPENDABILITY'S ATTRIBUTES OF DATA WAREHOUSE SYSTEM

A. Dependability's attributes

The original definition of dependability for a computing system is the ability to deliver service that can legitimately be trusted [13]. This definition induces the need for justification of trust to avoid service failure and make it more acceptable.

Dependability may be viewed according to different, but complementary attributes, which are summarized in: (i) Availability: ability of the system to be operational at the time requested (ii) Reliability: continuity of service, (iii) Security: non-occurrence of environmental catastrophic consequences, (iv) Confidentiality: the prohibition of unauthorized disclosure of information, (v) Integrity: non-occurrence of improper alteration of information, (vi) Maintainability: the ability to handle repairs and updates [13]. Associating integrity and availability together, as well as confidentiality, lead to security.

B. Data Warehouse System

DWS are known as a collection of decision support technologies, aiming at making decisions better and faster. ETL (Extraction, Transformation, Loading) retrieves data from various operational databases, cleans, transforms and loads it in DW repository [2]. The rules used to clean and transform the data and perform the transformation are part of the metadata management system. Users can extract data from the DW using, for instance, query tools. The components of a typical DWS are shown in "Fig.1".

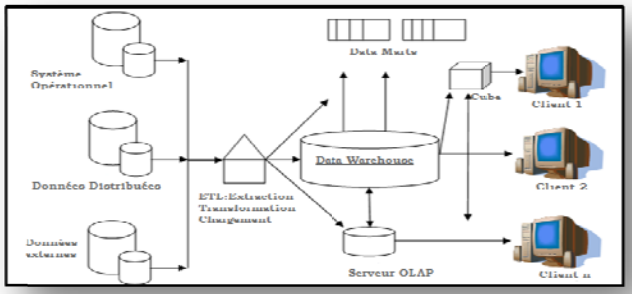


Figure1.Data Warehouse System's architecture

C. DWS's dependability

In this section we will adapt each dependability attribute in the DWS context:

1) Availability

Availability is the readiness of the system to deliver correct service whenever solicited [13]. The same definition is projectable on DWS. Thus, we can define the availability of the DWS as the latter's ability to provide extracted, cleaned, rebuilt information whenever needed, and which can be analyzable following different axes.

Literature provides many optimization techniques to deal with the demand of availability in DWS. Most of these techniques focus on DW, and are inherited from traditional relational database systems. Among these techniques [14] we find: materialized views, indexing methods, data partitioning [15], and parallel processing [16]. Regarding ETL, many techniques have been developed to improve its availability as it's mainly responsible for transforming and delivering data. Those techniques have proposed to optimize the ETL process, using the workflow concepts [17] and the concept of near-real-time data warehousing [18, 19].

2) Reliability

The common Reliability's definition consists of the continuity of correct service [13]. Projecting this definition to DWS allows us to consider this attribute as the percentage of time the DWS is available for use. While considering aspects of maturity, fault tolerance and recoverability.

Considering the definitions of availability and reliability, both emphasize the avoidance of failures, and can be grouped together, and collectively defined as the avoidance or minimization of service outages.

3) Maintainability

Maintainability is the ability to bear repairs and updates [13]. In reality, DWS are interacting with a dynamic environment due to changing business rules or operational objectives on one hand, and increasing requirements and changing in its data sources on the other hand. Thus, their maintenance is very important to insure their ability of adaption to their environment's evolutions.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Based on the previous work, we have to distinguish between two concepts:

- Maintainability: which is the ability of an element to be maintained; this ability stems from all design features that enhance the ability of service.
- Maintenance: is a series of actions of appropriate character (content, timing, quality) to restore or keep an element in an operational state.

Based on the definitions above, the two concepts are not contradictory, since maintenance is the action of maintaining. More the system accepts maintenance's operation, the more it's maintainable. In our proposal, we consider that the two aspects have the same meaning.

4) Security

Security is a basic requirement for diverse kinds of applications. Since business information in the data warehouse is sensitive, the different aspects (availability, integrity and confidentiality) of security are particularly important. In this way, information security should be taken into consideration from the early stages of DWS's development life-cycle [9, 10].

To summarize, the integration of dependability aspects is a necessary to insure a dependable DWS. But the challenge in this context is that the attributes are often: (i) Heterogeneous; (ii) Ranked differently and subjectively by the stakeholders; (iii) Interrelated: in conflict or in harmony. In addition, the implementation of these attributes requires a lot of expertise and domain knowledge.

IV. MDA : MODEL DRIVEN ARCHITECTURE

MDA [20] provides model-driven software development based on the separation of the specification of the system functionality and its implementation. It defines models at different abstraction levels, and transformations required to the passage of a level to another [21] "Fig. 2".

This approach propose business models (CIM: Computation Independent Model) based on system requirements. Conceptual models (PIM: Platform Independent Model), which do not include information about specific platforms and technologies. These latter are including in the logical models (PSM: Platform Specific Model). Developing information systems applying the MDA approach improves productivity, saving time and effort, and provides support for system evolution, integration, interoperability, portability, adaptability and reusability [22].

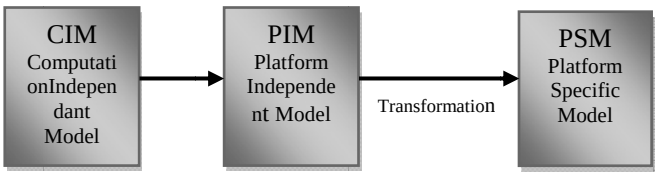


Figure 2. Generic MDA Models

The requirements of the future system are described in the CIM, which is refined into the PIM. The PSM is the result of the PIM transformation. The process refines the PIM based on the specification described in the Platform Description Module (PDM) defining how to use a specific platform.

The main advantages [21] of the MDA framework are: (i) Results are automatically generated which is expected to improve productivity, development duration, and cost; (ii) The developer pay more attention to CIM and PIM instead of focus on logical and technical details; (iii) PIM is portable to different target platforms; (iv) Once a transformation has been developed, it can be reused whenever needed; (v) Changes have to be done in the PIM only if the target platform has changed; (vi) New requirements in the CIM are passed from PIM to PSM immediately and changes are reflected automatically [14], [15], [16].

V. OUR APPROACH

Our proposal aims to provide a dependable DWS process design, taking into account, since the early stages of modeling, dependability's constraints. As these latter vary according to the field of use of DWS, we suggest a generic approach aligned to the standard MDA, allowing their integration in a refined and incremental way. Of course it based on different levels of abstraction to ensure traceability, and check their compatibility while considering their interaction. "Fig. 3" explicitly represents our approach.

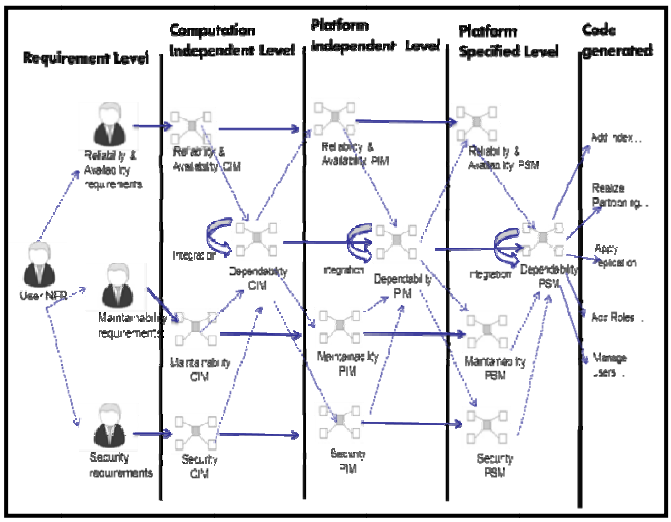


Figure 3. Our proposed approach

Our approach covers the entire DWS project. However, the DWS is itself a heterogeneous system, because of the platforms from which it is composed (DS, ETL, DW, analysis tools), we adapted the "divide to conquer" strategy. On one hand, it allows to deal with each layer preserving its own characteristics, and secondly integrate them with the surrounding layers. We have thus split the standard architecture of DWS into four homogeneous layers: the DS, ETL, DW and analysis tools. Then we conducted a qualitative study to list all parameters influencing the aspects

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

of dependability. We have integrated them through building a structured approach respecting the different types of models proposed by the MDA approach.

A. Requirement Level

This section is devoted to the qualitative study of all dependability's attributes in the four layers of the DWS. We isolated each of its layers. Based on the fact that the availability and reliability are very close (see section 3). They will therefore be associated in the following of our work. We can list the parameters that influence dependability's attributes considering that their implementation depends on the type of their attribution.

1) Data Source (DS):

Availability and Reliability <Structuration, Volumetric, Optimization, Data Quality, Support, Replication> as:

- •Structuration <structured, mi structured, not structured>
- •Volumetric <transaction frequency, recording number >
- •Optimization <index, fragmentation, view, parallelism >
- •Data Quality <redundant, missing, incomplete, mistaken>
- •Support <hard disc, communication>

Maintainability: <Documentation, Structuration, Optimization, Support> as:

- •Documentation <schema, metadata>
- •Structuration <structured, mi structured, not structured>
- •Optimization <index, fragmentation, view, parallelism >
- •Support <hard disc, communication>

Security: <User Management, Data, Support> as:

- •User Management< authentification, permission, quotas, audit, hierarchy>
- •Data <integrity, authenticity, completeness, locking>
- •Support <firewall, encryption, replication>

2) Extraction, Transformation, Loading (ETL)

Availability and Reliability <Data Quality, Periodicity, Cleaning, Complexity, Support> as

- •Data Quality <redundant, missing, incomplete, mistaken>
- •Periodicity <batch, real time, cycle refreshment>
- •Cleaning <consolidation, formatting, restructuration, normalization, doubloon elimination, aberrant values elimination, missing values >
- •Complexity < data size, calculation rules, DS number>
- •Support <RAM, temporary storage, communication>

Maintainability: <Metadata, Modularity, Complexity, Restoration> as:

- •Metadata <source, destination, cleaning operations, management rules, constraints>
- •Modularity <interdependence transformations, subsystems decomposition >
- •Complexity <calculation rules, DS number>
- •Restoration <recovery management, exception management>

Security: <Intrusion Management, Audit, Restoration> as:

- •Intrusion Management <firewall, encryption, codification, access control >
- •Audit <data, transformation>
- •Restoration <recovery management, exception management>

3) Data Warehouse (DW) Repository

Availability and Reliability <Schema, Volumetric, Complexity, Replication, Optimization, Support, loading Frequency> as:

- •Schema <star, snowflake, constellation>
- •Volumetric <recording number, access frequency, granularity>
- •Complexity < fact number, dimension number, hierarchy number>
- •Optimization <index, fragmentation, view, parallelism>
- •Support <hard disc, communication>

Maintainability <Documentation, Architecture, Complexity, Optimization, Support> as:

- •Documentation <metadata, schema, dimension>
- •Architecture <centralized, distributed, consolidation Data Mart>
- •Complexity <dependence dimensions, granularity>
- •Optimization <index, fragmentation, view, parallelism>
- •Support <hard disc, warm maintenance, cold maintenance >

Security <User Management, Data Management, Support> as:

- •User Management <authentification, permission, quotas, audit, hierarchy>
- •Data Management <integrity, authenticity, completeness, locking>
- •Support <Firewall, encryption, replication>

4) Restitution tools

Availability and Reliability <Type, Volumetric, Complexity, Support> as:

- •Type < Ad hoc query, parameterizable query, dashboard, cube OLAP, Data Mining>
- •Volumetric <data size>
- •Complexity <axis number, transformation, granularity>
- •Support <RAM, Communication>

Maintainability <Type, Documentation, Modularity, Complexity> as:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

•Type < Ad hoc query, parameterizable query, dashboard, cube OLAP, Data Mining>

•Documentation <metadata, applications, indicators, calculation rules>

•Modularity <interdependence, parallelism>

•Complexity <Axis number, transformation, granularity>

Security: < User Management, Data Management, Support> as:

•User Management<authentification, permission, quotas, audit, hierarchy>

•Data Management <integrity, validity, privacy>

•Support <firewall, encryption>

B. Computation Independent Level

At this stage, we will implement the Computation Independent Model (CIM) using the NFR (Non Functional Requirements) Framework approach. This model is based on the parameters from the previous level. The NFR Framework proposed by Chung [23], is based on the concept of "Softgoal" to represent information about the different qualities of a system, and their interdependencies [24].

The development of interactive SIG (Softgoal Interdependency Graphs) allows representing the dependability's attributes and their integration. Then, it can examine the interactions and identify potential conflicts, to better suggest a compromise in the early stages of the DWS project. The following diagrams "Fig. 4, 5, 6" clarify the CIM availability & reliability, security and maintainability of DWS.

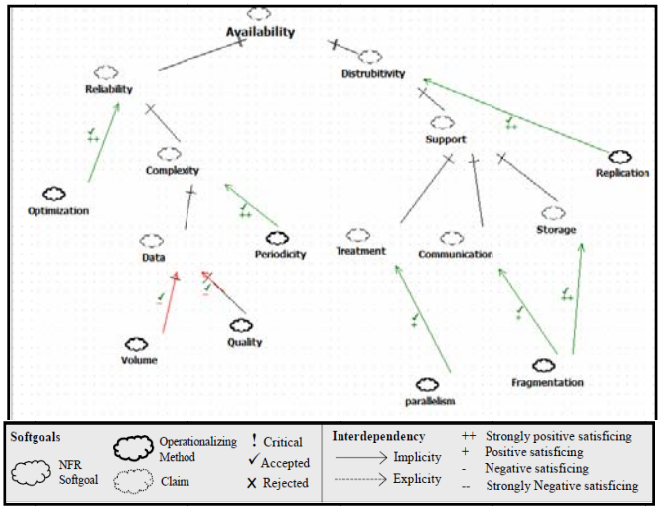


Figure 4. Availability & Reliability CIM

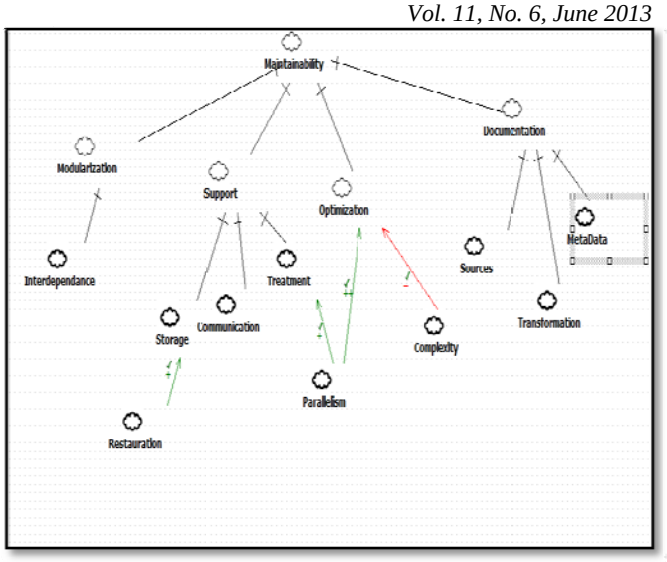


Figure 5. Maintainability CIM

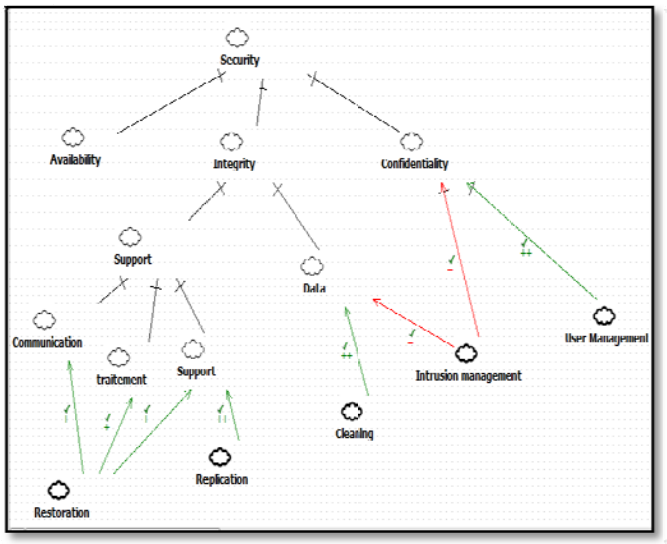


Figure 6. Security CIM

"Fig.7" shows an example of integration of the above models ("Fig. 4, 5, 6"). This integration will allow the designer to view all the dependability's constraints in a model. Therefore, the appropriate implementation's decisions of functional modules will be made taking into account non-functional requirements relating to dependability's one.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

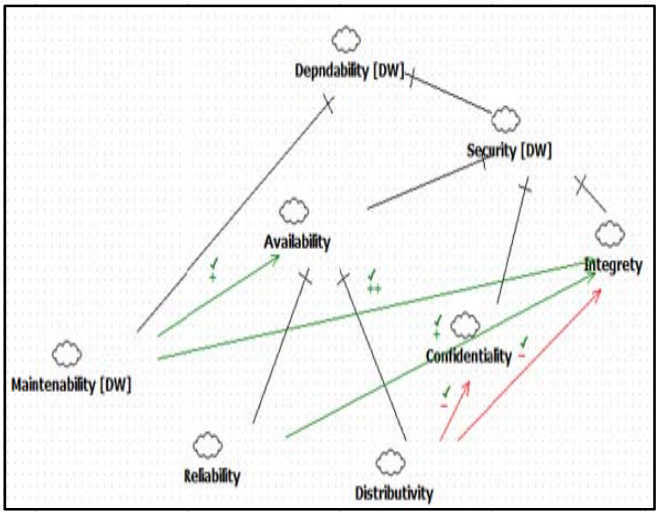


Figure7. Example of CIM integration

C. Platform Independent Level

1) The transition from CIM to PIM

To ensure traceability of dependability's aspects since the expression of requirements to the implementation, we was based on the work of S. Supakkul, L.Chung [25]. The authors proposed an integrated modeling language by extending UML with the NFR framework using the UML profile. They defined a MetaModel to represent the concepts in the NFR Framework. They alsoidentified the extension points for associating the two notations.

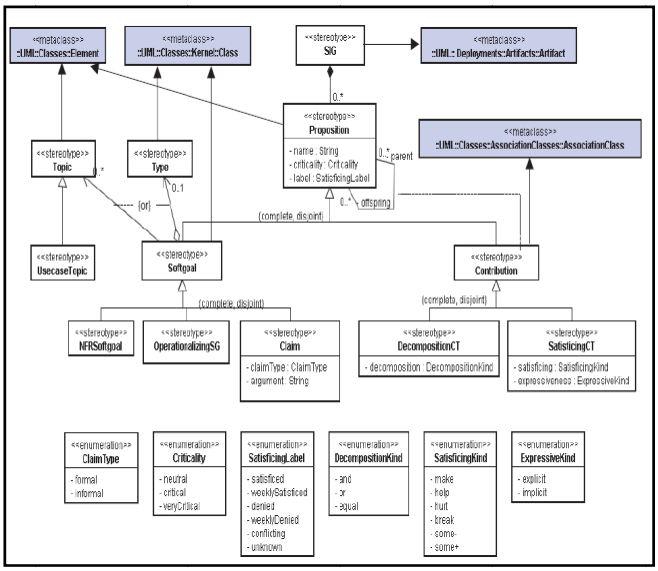


Figure 8. MetaModel of NFR Framework & relationship with UML metadata from [25]

"Fig. 8" explains the relationship between the Meta model of NFR Framework and the UML MetaModel. Therefore, we justified the existence ofthe transformation from our CIM (using NFR Framework) to PIM (using UML profile) by exploiting this relationship [26].

2) Development of PIM

In this part we relied on an existing UML profile [27]: UML profile for Quality of Service "QoS" to represent the properties of quality of service that include dependability. "Fig. 9, 10, 11"represent the PIM of availability & reliability, maintainability and security according the profile already cited "Fig.12" shows an example of their integration.

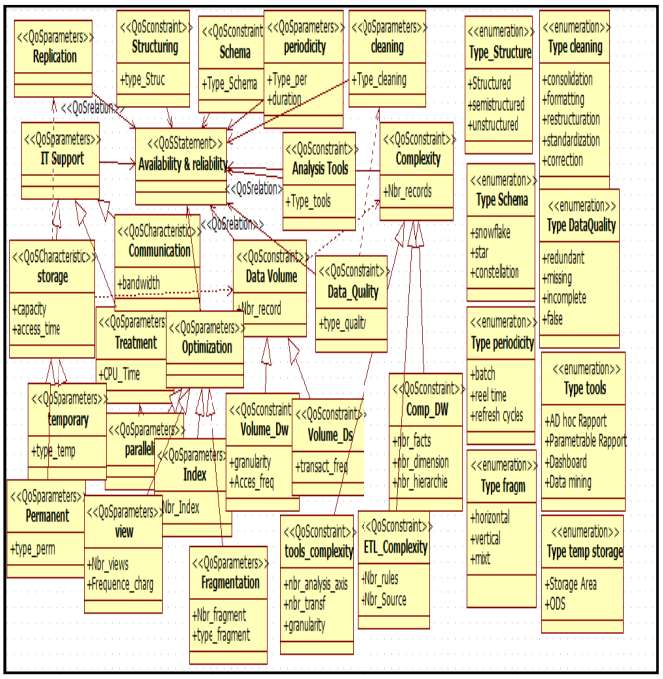


Figure 9. PIM's Availability & reliability

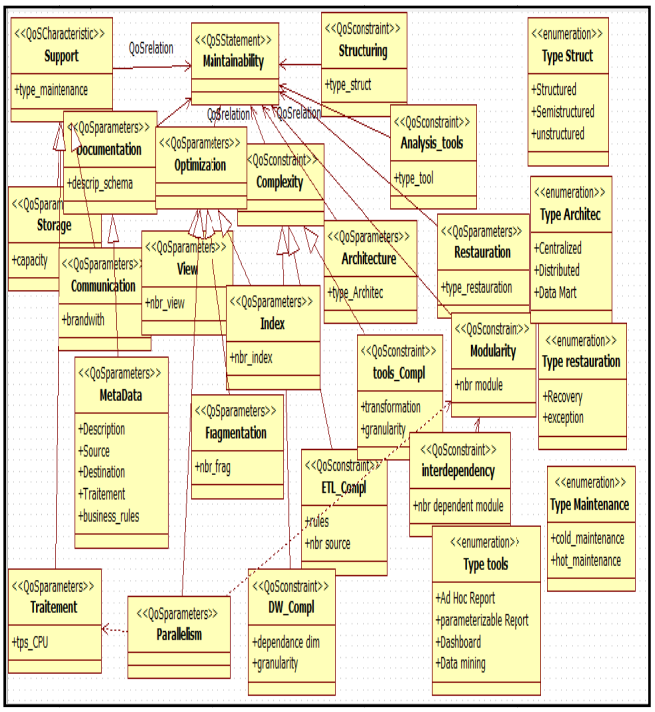


Figure 10. Maintainability PIM

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

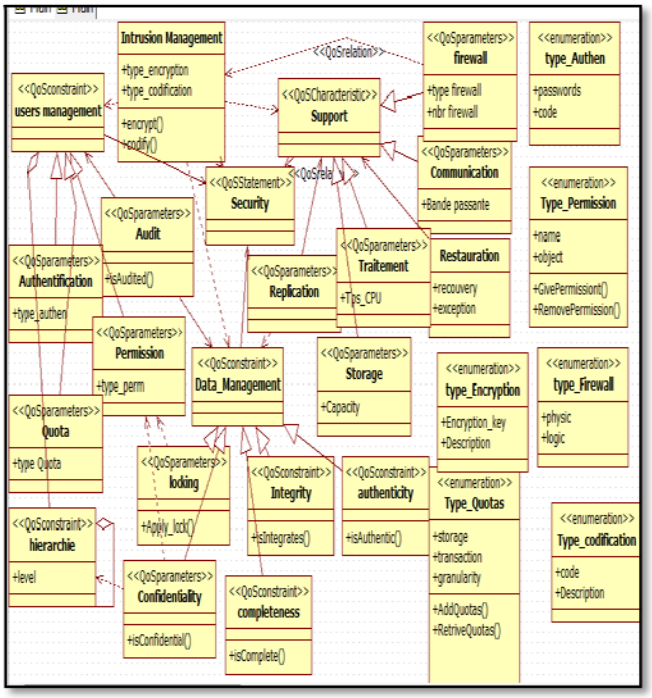


Figure 11. Security PIM

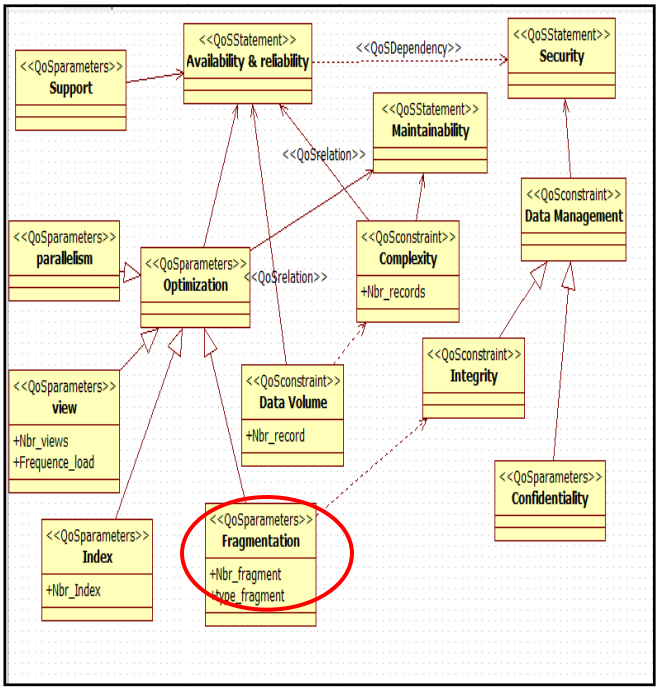


Figure12. PIM which integrate the dependability's attributes

D. Specified Level Platform

At this level, developers need to combine information on the PDM platform (Platform Description Model) with PIM to generate the PSM. Given the heterogeneity of platforms which composed DWS architecture, we suppose that specific PSM must be developed for each of dependability attributes.

However the DWS is a system composed of several operatively different layers. Thus, for each layer of the DWS and every aspect of the dependability, a PSM must be establishedto keep the specifications of each layer. To keep our approach valid for all platforms, we will do not specify the PSM because they usea specific technology.

Once the PSM are refined, the generation code is automatic. In addition, the integration of dependability's attributes with the functional part is guaranteed through the aspect oriented programming. In the next section we present a case study to demonstrate the validity of our approach.

VI. EXAMPLE OF IMPLEMENTATION

In our case study we take example of implementation of some of aspects of the Business Intelligence dependability, where users require as priority needs, data availability and reliability. However, it supposed that data are often maintained. Thus to ensure a dependable DWS we must take into account the interaction between the three aspects.

Based on CIM and PIM proposed by our approach, we were able to select the common parameters to the various attributes that are availability, reliability and maintainability, and on which we can act. These parameters are: replication, support and optimization (index, views, fragmentation, parallelism).

Support and replication technics can be directly applied on the media platform of DWS. We chose to implement fragmentation in the DW repository. The tests were performed by a workload query running on the APB-1 benchmark [28] under Oracle 10g. We applied the horizontal fragmentation of 10 queries Q1…Q10 of thebenchmark. The response time comparison of these queries before and after fragmentation is illustrated by "Fig. 13".

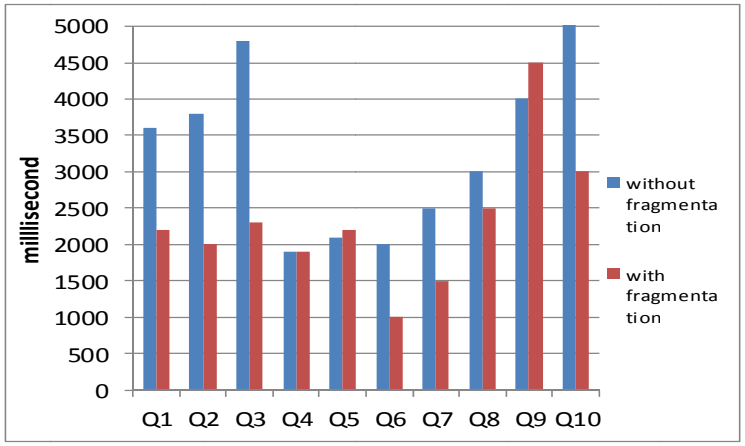


Figure 13. Graphical representation of results

In the following the summary of results:

- ✓ Fragmentation has improved the response time of the data warehouse which has a positive impact on the availability and reliability.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

- ✓ Operations of maintainability and security can be slowed because of the complexity generated by the different fragments.
- ✓ We can adjust the number of fragments in order to find a compromise between availability, reliability and maintainability, security taking into account the levels tolerated by users.

VII. CONCLUSION AND PERSPECTIVE

Our approach offers models at different levels of abstraction to deal with attributes of dependability. The main difficulty was to select measurable parameters common to the various aspects of the dependability. This problem has been overcome by our proposal, which integrates the different attributes taking into account their interaction from the early phases of modeling. To realize this, we used the models offered by the MDA.

As perspective, we can propose the integration of aspect-oriented programming approach in order to concretize the implementation of our proposal. We can also suggest the definition of an UML profile specific to dependability in DWS's context.

REFERENCES

[1] R. Kimball,L.Reeves,M.Ross, andW.Thornthwaite, The Data Warehouse Lifecycle Toolkit, Wiley, 2004

[2] W.Inmon, Building the Data Warehouse, 3rd edition, Wiley, 2002.

[3] A. Avizienis, J. Laprie, and B. Randell, "Fundamental concepts of dependability," in Proc. 3rd Information Survivability Workshop, 2000, pp. 7–12.

[4] M.Golfarelli,: From user requirements to conceptual design in data warehouse design: A survey. In: Bellatreche, L. (ed.) Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction. IGI Global (2009)

[5] F. R. S Paim, J. F. B Castro, (2003, September). DWARF: An approach for requirements definition and management of data warehouse systems. InRequirements Engineering Conference, 2003. Proceedings. 11th IEEE International (pp. 75-84). IEEE.

[6] F. R. S Paim, J. Castro, (2002, November). "Enhancing Data Warehouse Design with the NFR Framework". In *5th Workshop on Requirements Engineering (WER2002), Valencia, Spain* (pp. 40-57).

[7] V Peralta,A. Illarze, &R. Ruggia, (2003). On the applicability of rules to automate data warehouse logical design.In *CAiSE Workshops*.

[8] E.Soler,V. Stefanov,J. N. Mazón,J.Trujillo, ,E. Fernández-Medina, , &M.Piattini, (2008, March). Towards comprehensive requirement analysis for data warehouses: Considering security requirements. . In: The Third International Conference on Availability, Reliability and Security ARES 2008, Barcelone, Espagne, pp. 104–111. IEEE Computer Society, Los Alamitos (2008)

[9] E.Soler, R.Villarroel,J. Trujillo, E.Fernandez-Medina, ,M. &Piattini, (2006, April). Representing security and audit rules for data warehouses at the logical level by using the common warehouse metamodel. In Availability, Reliability and Security, 2006. ARES 2006. The First International Conference on (pp. 8-pp). IEEE.

[10] C.Blanco, I.Garcia-Rodriguez de Guzman, E.Fernandez-Medina, , & J.Trujillo, (2010). Defining and transforming security rules in an MDA approach for DWs. International Journal of Business Intelligence and Data Mining, 5(2), 116-133.

[11] L. Bellatreche, (2009). Data Warehousing Design and Advanced Engineering Applications: Methods for Complex Construction (Advances in Data Warehousing and Mining-Adwm. Information Science Reference-Imprint of: IGI Publishing.

[12] E. Lau,&S.Madden, (2006, September). An integratedapproach to recovery and highavailability in an updatable, distributed data warehouse. In*Proceedings of the 32nd international conference on Very large data bases* (pp. 703-714). VLDB Endowment.

[13] AlgirdasAvizienis, Jean-Claude Laprie, Brian Randell, and Carl Landwehr,Basic Concepts and Taxonomy of Dependable and Secure ComputingIEEE Transactions On Dependable And Secure Computing, Vol. 1, No. 1, January-March 2004

[14] S.Chaudhuri,U.Dayal,&V. Narasayya, (2011). An overview of business intelligence technology. Communications of the ACM, 54(8), 88-98

[15] L.Bellatreche, M.Schneider, H. Lorinquer, &M.Mohania, (2004). Bringing together partitioning, materialized views and indexes to optimize performance of relational data warehouses. In Data Warehousing and Knowledge Discovery (pp. 15-25).Springer Berlin Heidelberg.

[16] F.Dehne, T.Eavis, & A.Rau-Chaplin, (2003, May). Parallel multi-dimensional ROLAP indexing.In Cluster Computing and the Grid, 2003.Proceedings.CCGrid 2003. 3rd IEEE/ACM International Symposium on (pp. 86-93). IEEE.

[17] A.Simitsis,P. Vassiliadis, &T.Sellis, (2005, April). Optimizing ETL processes in data warehouses. In Data Engineering, 2005.ICDE 2005.Proceedings. 21st International Conference on (pp. 564-575). IEEE.

[18] T.Jörg,&S.Dessloch, (2010). Near real-time data warehousing using state-of-the-art ETL tools.In Enabling Real-Time Business Intelligence (pp. 100-117).Springer Berlin Heidelberg.

[19] R. J.Santos, & J.Bernardino, (2008, September). Real-time data warehouse loading methodology. In Proceedings of the 2008 international symposium on Database engineering & applications (pp. 49-58).ACM.

[20] D. S. Frankel, (2003). *Model Driven Architecture ApplyingMda*. John Wiley& Sons.

[21] A. G.Kleppe,J. B. Warmer,&W.Bast, (2003). *MDA explained, the model driven architecture: Practice and promise*. Addison-Wesley Professional.

[22] C.Blanco, I. Garcia-Rodriguez de Guzman, E. Fernandez-Medina, & J. Trujillo. (2010). Defining and transformingsecurityrules in an MDA approach for DWs. *International Journal of Business Intelligence and Data Mining, 5*(2), 116-133.

[23] N. Subramanian, L. Chung, (2001, September). Software architecture adaptability: an NFR approach. In *Proceedings of the 4th International Workshop on Principles of Software Evolution* (pp. 52-61). ACM.

[24] L. Chung, , J. C. S. do Prado Leite, (2009). On non-functional requirements in software engineering. In *Conceptual modeling: Foundations and applications*(pp. 363-379). Springer Berlin Heidelberg.

[25] S. Supakkul, L.Chung, (2005, August). A UML profile for goal-oriented and use case-drivenrepresentation of NFRs and FRs. In *Software Engineering Research, Management and Applications, 2005. Third ACIS International Conference on* (pp. 112-119). IEEE.

[26] M. R.Abid,D. Amyot, ,S. S. Somé, &G.Mussbacher, (2009). *A UML profile for goal-orientedmodeling* (pp. 133-148). Springer Berlin Heidelberg.

[27] J. Ø.Aagedal, &E. F.Ecklund Jr, (2002). ModellingQoS: towardsa UML profile. In *≪ UML≫ 2002—The UnifiedModelingLanguage* (pp. 275-289). Springer Berlin Heidelberg.

[28] Council, O. L. A. P. (1998). Apb-1 olap benchmark, release ii. *http://www.olapcouncil. org/research/bmarkly.htm*.

A Novel Framework using Similar to Different Learning Strategy

Bhanu Prakash Battula
Research scholar,
Acharya Nagarjuna University,
Guntur, India.
.

Dr. R. Satya Prasad
Associate Professor,
Acharya Nagarjuna University,
Guntur, India.
.

Abstract— **Most of the existing classification techniques concentrate on learning the datasets as a single similar unit, in spite of so many differentiating attributes and complexities involved. However, traditional classification techniques, require to analysis the dataset prior to learning and for not doing so they loss their performance in terms of accuracy and AUC. To this end, many of the machine learning problems can be very easily solved just by careful observing human learning and training nature and then mimic the same in the machine learning.**

In response to these issues, we present a comprehensive suite of experiments carefully designed to provide conclusive, reliable, and significant results on the problem of efficient learning. This paper proposes a novel, simple and effective machine learning paradigm that explicitly exploits this important similar-to-different (S2D) human learning strategy, and implement it based on three algorithms (C4.5, CART and LR) efficiently. The framework not only analyzes the datasets prior to implementation, but also carefully allows classifier to have a systematic study so as to mimic the human training technique designed for efficient learning. Experimental results show that the method outperforms the state of art methods in terms of learning capability and breaks through the gap between human and machine learning. In fact, the proposed method similar-to-different (S2D) strategy may also be useful in efficient learning of real-world complex and high dimensional datasets, especially which are very typical to learn with traditional classifiers.

Keywords- Data Mining, Classification; learning strategy;

Similar-to-Different (S2D).

I. INTRODUCTION

One of the research hotspots in the field of machine learning is classification. There are different types of classification models such as decision trees, SVM, neural networks, Bayesian belief networks, Genetic algorithm etc. The simple structure, the wide applicability on real time problems, the high efficiency and the high accuracy are the strengths for decision trees. In recent years, many authors proposed improvements in decision trees learning strategy [1,2]. The use of expert knowledge in pruning decision trees for applicability in medical analysis is proposed in [3]. A large number of classifiers build the model of dataset for classification by using the traditional learning strategies. On the other hand, the traditional learning techniques are bottle necked the performance of the datasets. However, several investigations also suggest that there are other factors that contribute to such performance degradation, for example, size of the dataset, density of the dataset, and overall complexity of the dataset. This work focuses on the analysis of improved learning strategy for the open problems related to complex and high dimensional dataset.

In this paper, we propose a novel, simple and effective machine learning paradigm that uses this important similar-to-different learning strategy, called *S2D* (Similar to Different). Its applications in human-oriented learning tasks, especially cognitive learning tasks, will be fruitful. S2D builds its model with similar examples first. More specifically, it selects those examples that are close to each other (thus similar), and updates the model with them. S2D works iteratively in this way which is almost the same as the human learning form of similar to different process. Experiment results show that our new learning paradigm S2D has several distinctive advantages over C4.5.

First of all, S2D does indeed take much less effort in building its model than C4.5. Second, minimal effort learning implies that the process of learning and the final learned model are more stable and reliable. This is certainly crucial for human learning, as well as for

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

machine learning applications. Finally, even though S2D only locally updates the model with minimal effort, we show that it is as accurate as the global learner C4.5. One might think that as S2D always takes the similar example to update its model locally and incrementally, it may not predict as accurately as the global learner C4.5 which builds its model on the whole dataset. Our experiment results show that S2D predicts only slightly worse than C4.5.

The rest of this paper is organized as follows. Several previous works related to different learning strategies are reviewed in Section 2. Section 3 describes a generic S2D paradigm. Section 4 presents the datasets and measures used for validation. Section 5 presents experimental settings for validation of S2D based on the decision tree learning algorithm. Experiment results are shown in Section 6. Section 7, conclude our work with future scope.

II. LITERATURE REVIEW

The cognitive validation of our Similar to Different strategic model of human iterative-learning process can be justified by the prominent works done in the cognitive science [28-32]. The other recent works in this area are, Serkan celik *et al.* [4] have examine vocabulary-learning strategies adopted by Turkish EFL students, specifically the frequencies and helpfulness ratings of strategy use, strategy patterns, as well as their change for students across different language levels. The study involved 95 tertiary level English as a foreign language learners. Data were analyzed statistically and the results indicated that the participants' general use of vocabulary learning strategies was somewhat inadequate and there was a gap between their use of strategies and related perceptions of strategy usefulness. Zhou Guo Dong *et al.* [5] have proposed a novel hierarchical learning strategy to deal with the data sparseness problem in semantic relation extraction by modeling the commonality among related classes. For each class in the hierarchy either manually predefined or automatically clustered, a discriminative function is determined in a top-down way. As the upper-level class normally has much more positive training examples than the lower-level class, the corresponding discriminative function can be determined more reliably and guide the discriminative function learning in the lower-level one more effectively, which otherwise might suffer from limited training data. Authors proposed, two classifier learning approaches, i.e. the simple perceptron algorithm and the state-of-the-art Support Vector Machines, are applied using the hierarchical learning strategy.

Edwin Lughofer [6] has proposed a novel active learning strategy for data-driven classifiers, which is based on unsupervised criterion during off-line training phase, followed by a supervised certainty-based criterion during incremental on-line training. In this sense, they call the new strategy hybrid active learning. Sample selection in the first phase is conducted from scratch (i.e.no initial labels/learners are needed) based on purely unsupervised criteria obtained from clusters: samples lying near cluster centers and near the borders of clusters are expected to represent the most informative ones regarding the distribution characteristics of the classes. In the second phase, the task is to update already trained classifiers during on-line mode with the most important samples in order to dynamically guide the classifier to more predictive power. Both strategies are essential for reducing the annotation and supervision effort of operators in off-line and on-line classification systems, as operators only have to label an exquisite subset of the off-line training data representation give feedback only on specific occasions during on-line phase.

Kevin Duh *et al.* [7] have proposed a flexible transfer learning strategy based on sample selection. Source domain training samples are selected if the functional relationship between features and labels do not deviate much from that of the target domain. This is achieved through a novel application of recent advances from density ratio estimation. The approach is flexible, scalable, and modular. It allows many existing supervised rankers to be adapted to the transfer learning setting. Xiaodong Yu *et al.* [8] have proposed a novel updating algorithm based on iterative learning strategy for delayed coking unit (DCU), which contains both continuous and discrete characteristics. Daily DCU operations under different conditions are modeled by a belief rule-base (BRB), which is then, updated using iterative learning methodology, based on a novel statistical utility for every belief rule. Compared with the other learning algorithms, their methodology can lead to a more optimal compact final BRB. With the help of this expert system, a feed forward compensation strategy is introduced to eliminate the disturbance caused by the drum-switching operations.

R.J. Gil *et al.* [9] have proposed a novel model of an Ontology-Learning Knowledge Support System (OLeKSS) is proposed to keep these KSSs updated. The proposal applies concepts and methodologies of system modeling as well as a wide selection of OL processes from heterogeneous knowledge sources (ontologies, texts, and databases), in order to improve

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

KSS's semantic product through a process of periodic knowledge updating. An application of a Systemic Methodology for OL (SMOL) in an academic Case Study illustrates the enhancement of the associated ontologies through process of population and enrichment.

Md Nasir *et al.* [10] have proposed a variant of single-objective PSO called Dynamic Neighborhood Learning Particle Swarm Optimizer (DNLPSO), which uses learning strategy whereby all other particles' historical best information is used to update a particle's velocity as in Comprehensive Learning Particle Swarm Optimizer (CLPSO). But in contrast to CLPSO, in DNLPSO, the exemplar particle is selected from a neighborhood. This strategy enables the learner particle to learn from the historical information of its neighborhood or sometimes from that of its own. Moreover, the neighborhoods are made dynamic in nature i.e. they are reformed after certain intervals. This helps the diversity of the swarm to be preserved in order to discourage premature convergence. Biao Niu *et al.* [11] have proposed a novel batch mode active learning scheme for informative sample selection. Inspired by the method of graph propagation, we not only take the correlation between labeled samples and unlabeled samples, but the correlation among unlabeled samples taken into account as well. Especially, considering the unbalanced distribution of samples and the personalized feedback of human we propose an asymmetric propagation scheme to unify the various criteria including uncertainty, diversity and density into batch mode active learning in relevance feedback.

Ching-Hung Lee *et al.* [12] have proposed a hybrid of algorithms for electromagnetism-like mechanisms (EM) and particle swarm optimization (PSO), called HEMPSO, for use in designing a functional-link-based Petri recurrent fuzzy neural system (FLPRFNS) for nonlinear system control. The FLPRFNS has a functional link-based orthogonal basis function fuzzy consequent and a Petri layer to eliminate the redundant fuzzy rule for each input calculation. In addition, the FLPRFNS is trained by the proposed hybrid algorithm. The main innovation is that the random-neighborhood local search is replaced by a PSO algorithm with an instant-update strategy for particle information. Each particle updates its information instantaneously and in this way receives the best current information. Thus, HEMPSO combines the advantages of multiple-agent-based searching, global optimization, and rapid convergence. Gwénolé Quellec *et al.* [13] have proposed a novel multiple-instance learning framework, for automated image classification. Given reference images marked by clinicians as relevant or irrelevant, the image classifier is trained to detect patterns, of arbitrary size, that only appear in relevant images. After training, similar patterns are sought in new images in order to classify them as either relevant or irrelevant images. Therefore, no manual segmentations are required. As a consequence, large image datasets are available for training.

Costantino Grana *et al.* [14] have proposed a novel algorithm to synthesize an optimal decision tree from OR-decision tables, an extension of standard decision tables, complete with the formal proof of optimality and computational cost analysis. As many problems which require recognizing particular patterns can be modeled with this formalism, They select two common binary image processing algorithms, namely connected components labeling and thinning, to show how these can be represented with decision tables, and the benefits of their implementation as optimal decision trees in terms of reduced memory accesses.

Table 1 Recent Advances in class imbalance learning

ALGORITHM	DESCRIPTION	REFERENECE
Hi-SVM Hi-PA	Hierarchical learning Strategy On SVM and Perceptron Algorithm.	[5]
IL-RIMMER	Iterative learning-RIMMER Algorithm.	[8]
OLeKSS	Ontology-Learning Knowledge Support System	[9]
DNLPSO	Dynamic Neighborhood Learning Particle Swarm Optimizer	[10]
APAL	Asymmetric Propagation based Active Learning algorithm	[11]
HEMPSO	Hybridization of ElectroMagnetism like and Particle Swarm Optimization Algorithm	[12]
MIL	Multiple Instance Learning Framework	[13]
MGDT	Maximum Gain Decision Tree for OR-Decision Tables	[14]
GCSDT	Genetically optimized Cluster Oriented Soft Decision Trees	[16]

Joel E. Denny *et al.* [15] have demonstrate that a well-known algorithm described by David Pager and implemented in Menhir, the most robust minimal LR(1) implementation they have discovered that, it does not always achieve the full power of canonical LR(1) when the given grammar is non-LR(1) coupled with a specification for resolving conflicts. They also detail an

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

original minimal LR(1) algorithm, IELR(1) (Inadequacy Elimination LR(1)), which they have implemented as an extension of GNU Bison and which does not exhibit this deficiency. Sanjay Kumar Shukla *et al.* [16] have developed a novel methodology, genetically optimized cluster oriented soft decision trees (GCSDT), to glean vital information imbedded in the large databases. In contrast to the standard C-fuzzy decision trees, where granules are developed through fuzzy (soft) clustering, in the proposed architecture granules are developed by means of genetically optimized soft clustering. In the GCSDT architecture, GA ameliorates the difficulty of choosing an initialization for the fuzzy clustering algorithm and always avoids degenerate partitions. This provides an effective means for the optimization of clustering criterion, where an objective function can be illustrated in terms of cluster's center. Growth of the GCSDT is realized by expanding nodes of the tree, characterized by the highest inconsistency index of the information granules.

Sanjay Jain *et al.* [18] have a present study aims at insights into the nature of incremental learning in the context of Gold's model of identification in the limit. With a focus on natural requirements such as consistency and conservativeness, incremental learning is analyzed both for learning from positive examples and for learning from positive and negative examples. In [19] authors introduced a novel form of decision tables, namely OR-Decision Tables, which allow including the representation of equivalent actions for a single rule. An heuristic to derive a decision tree for such decision tables was given, without guarantees on how good the derived tree was. In [20], authors presented a preliminary version of a bottom-up dynamic programming proposed by Schumacher et al. [21] which guarantees to find the optimal decision tree given an expanded limited entry (binary) decision table, in which each row contains only one non zero value. The summary of all the advances is presented in Table 1.

III. SIMILAR TO DIFFEREN T (S2D) LEARNING STRATEGY

Our novel similar-to-different learning strategy models the human iterative-learning process in which the similar set of materials is repeatedly studied, and for each iteration only the similar cases are learned.

The rationale behind S2D is that learning is a gradual process of accumulating knowledge. Learning similar set of examples first may make the learning of different examples easier, thus the whole learning process becomes easier and more reliable. The C4.5 [22] algorithm is a global learning algorithm i.e. it learns from all the dataset to generate the model and it selects instances randomly to build the model. This peculiar behavior of C4.5 leads to a limitation, where we can make a proper strategy to make the learning more effective. The best strategy for any machine is only a natural strategy, which humans follow for learning new things. The human strategy of learning new things is learning set of similar components and then going for another set of similar components which are different from earlier set and then learning recursively in the above fashion. We can name this strategy as similar to different (S2D) learning strategy. This strategy may also be known as curriculum learning strategy, since all academic curriculums will follow the same strategy.

At a high level, S2D is very simple, and it can use any classifier that can produce refined class probability estimation as its base learner. Generally speaking, for each iteration, S2D selects the similar example based on the similar sets and updates the model locally with the selected example, until all examples have been learned. However, several important issues deserve further explanations. First, how can S2D select the similar example? Second, how can S2D select the first similar example before any model is built? Third, how can S2D select the similar example when tie happens?

The first issue, how to select the similar example, is crucial for the S2D learning paradigm. Without a proper measurement, S2D cannot choose the similar example correctly. Here we propose a simple and effective measurement, which groups the instances with important attributes selected in filter. The nearer the range for that attribute, the similar the example is for the current model to learn. Thus a base learner that can generate refined class probability estimation is a requirement of S2D, as we mentioned before. Also the measurement is consistent with human intuition about similarity: the less the surprise (i.e., difference or error), the similar it is.

The second issue of selecting the first similar example can be tricky, as the current model is empty. We design a simple and effective strategy for S2D. S2D scans over all the similar set of training examples formed by correlation based feature subset filter to pick up the most similar example. If more number of similar subsets appears then an example from the majority subset will be chosen randomly.

The third issue, the tie-breaking strategy, can be crucial, if tie happens often when S2D selects the similar example with the current model. If ties happen, S2D must choose one randomly. This may affect the performance of S2D. Indeed, for some algorithms, such

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

as C4.5 (the base learner for S2D in this paper), ties do happen often. This is because C4.5 predicts all examples falling in the same leaf with the same prediction (i.e., same label and same probability estimation). For those algorithms, effective tie-breaking strategies are necessary. For C4.5, we use the number of positive and negative examples in each node as heuristic information for the tie breaking.
We implemented this S2D strategy in C4.5 learning process. We modified the learning process of C4.5 to suit with our S2D strategy. In the first phase, C4.5 is allowed to learn on similar set of instances in the dataset, the similarity in the instances can be measured by using metrics relating attributes of the instances. In the next phase of the approach, we need to expose C4.5 algorithm to learn on to the other similar set of instances, which are different from the earlier sets. This new learning strategy is applied to a base algorithm; in this case C4.5 is used to obtain measures such as accuracy, tree size, AUC, error rate, precision, recall and f-measure.

The algorithm for S2D approach is given below,

(Algorithm 1: Similar2Different Framework)

Input: Tr: a training sample set
La: Label for Tr
M: The number of selected Features
Output: Measures: Accuracy, Tree size, AUC, Error rate

Procedure:

```
Phase I: Finding important features using Filter
Features = Apply Filter (Dataset, CFS)

Phase II: Building Similar instance sets

for feature i = 1 to M  // M represents # features selected
for Tr (j) = 1 to N     //  N represents # similar subsets in that feature
Build Subset Tr (i.j)
end for
end for

 Phase III: Training on similar sets

For Tr (i,j) =1 to O // O represents # similar subset of each feature
LearnModel = Build(Tr (i.j), Base Algorithm)
End for
Predict (Learning Mode, Measure)
```

The algorithm 1: S2D can be explained as follows:

The inputs to the algorithm are training sample set "Tr" and label of training sample set "La". The output of the algorithm will be the average measures such as accuracy, tree size, AUC, error rate, precision, recall and f-measure produced by the S2D method. The algorithm begins with the initial stage of identifying important features M, where M is the number of features extracted by applying correlation-based feature subset filter [23] on the dataset. The 'M' value will change from one dataset to other, and depending upon the unique properties of the dataset the number of features can be more or less. In the next stage similar subsets are formed and consecutively trained on the base algorithm.

The different components of our new proposed framework are elaborated in the next subsections.

A. Finding important features using filter

How to find the most influencing attribute is the crucial issue. The important features are selected by using an attribute selecting filter; in this case we have used CFS i.e correlation based feature subset evaluation [23].

CFS can be defined as: A good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other [23]. The correlation based feature selection algorithm is a subset evaluation heuristic that takes into account the usefulness of individual features for predicting the class along with the level of inter-correlation among them. This heuristic is given below,
This heuristic is given as in equation 1,

$$Merit_s = \frac{kr_{cf}}{\sqrt{k + k(k-1)r_{ff}}} \qquad (1)$$

where, the heuristic "$Merit_s$" of a feature subset s containing k features, r_{cf} the average feature-class correlation, and r_{ff} the average feature-feature inter-correlation. The numerator can be thought of as giving an indication of how predictive a group of features are; the denominator of how much redundancy there is among them. The heuristic handles irrelevant features as they will be poor predictors of class. Redundant attributes are discriminated against as they will be highly correlated with one or more of the other features. The percentage of the weak instances removed may depend upon the properties of the dataset. The number of most influencing features in the dataset may also depend upon the unique properties of the dataset.

B. Building Similar Instances Sets

The building of similar instance sets depends upon the identifying of important attributes. If the attributes identified are less in number and the similar ranges are few then we can form small number of similar subsets for learning. If the attributes identified are more and the similar ranges in those attribute are more then similar subsets formed will be more. The selection of attributes and

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

forming of similar subsets will differ from dataset to dataset.

C. Training Model on Similar Sets

The similar examples for each subset are selected and trained to build the model in an iterative process. In this case we have considered C4.5 as the base algorithm. Voting and boosting methods are used to build the overall model. The metrics for evaluation such as accuracy, tree size, AUC, error rate, precision, recall and f-measure are measured.

IV. DATASETS AND MEASURES

We considered six benchmark real-world imbalanced dataset from the UCI machine learning repository [24] to validate our proposed method.

For every data set, we perform a tenfold stratified cross validation. Within each fold, the classification method is repeated ten times considering that the sampling of subsets introduces randomness. The accuracy, tree size, AUC, error rate, precision, recall and f-measure of this cross-validation process are averaged from these ten runs. The whole cross-validation process is repeated for ten times, and the final values from this method are the averages of these ten cross-validation runs.

Table 2 contains the name of the dataset, the total number of examples (Instances), attribute, the number of target classes for each dataset and the missing values.

Table 2 Summary of benchmark imbalanced datasets

Datasets	Instances	Attributes	Class	Missing
Arrhythmia	452	280	16	Y
Credit-g	1,000	21	2	N
Glass	214	10	7	N
Hepatitis	155	20	2	Y
Ionosphere	351	35	2	N
Waveform	5,000	41	3	N

Evaluation Criteria:

In this paper, we use accuracy, tree size, AUC, error rate, precision, recall and f-measure as performance evaluation measures.
Let us define a few well known and widely used measures [29]:
The most commonly used empirical measure; accuracy is computed by using the below equation (2),

$$ACC = \frac{TP + TN}{TP + FN + FP + FN} \tag{2}$$

Another measure for performance evaluation is AUC. A quantitative representation of a ROC curve is the area under it, which is known as AUC. When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TP rate and TN rate.

The Area under Curve (AUC) [33] measure is computed by equation (3),

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \tag{3}$$

The Precision [34] measure is computed by using eq (4),

$$\Pr ecision = \frac{TP}{(TP) + (FP)} \tag{4}$$

The F-measure [34] Value is computed by using eq (5),

$$F - measure = \frac{2 \times \Pr ecision \times \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call} \tag{5}$$

The Recall [34] measure is computed by using eq (6),

$$\mathrm{Re} call = \frac{TP}{(TP) + (FN)} \tag{6}$$

In these experiments, the size of the tree is calculated by the depth of the tree using number of nodes and leaves. Testing errors is computed as the number of errors produced when separate training and testing set is used for training and testing.

V. EXPERMENTAL SETTINGS

We performed the implementation using Weka [25] on Windows XP with 2Duo CPU running on 3.16 GHz PC with 3.25 GB RAM.

Algorithms and Parameters:

In first place, we need to define a baseline classifier which we use for our proposed learning strategy. With this goal, we have used C4.5 decision tree generating algorithm [23] and CART. C4.5 learning algorithm constructs the decision tree top-down by the usage of the normalized information gain (difference in entropy)

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is the one used to make the decision.

Algorithms for comparison:

To validate the proposed S2D algorithm, we compared it with the traditional C4.5, CART (Classification and Regression trees), FT (Functional Trees), REP (Reduced Error Pruning Tree), SMOTE (Synthetic Minority Oversampling TEchnique) and Logistic Regression (LR).

Specifically, we consider five different algorithmic approaches for comparison:

- C4.5: we have selected the C4.5 algorithm as a well-known classifier that has been widely used for imbalanced data. A decision tree consists of internal nodes that specify tests on individual input variables or attributes that split the data into smaller subsets, and a series of leaf nodes assigning a class to each of the observations in the resulting segments. For our study, we chose the popular decision tree classifier C4.5, which builds decision trees using the concept of information entropy. For this experimental set of C4.5 we have used all the default parameters in WEKA workbench.

- CART: The CART methodology is technically known as binary recursive partitioning. The process is binary because parent nodes are always split into exactly two child nodes and recursive because the process can be repeated by treating each child node as a parent. For this experimental set of CART [27], we have used all the default parameters in WEKA workbench. The key elements of a CART analysis are a set of rules for:

  i. Splitting each node in a tree;
  ii. Deciding when a tree is complete; and
  iii. Assigning each terminal node to a class outcome (or predicted value for regression).

- FT: Functional Trees (FT) is a classifier for building 'Functional trees', which are classification trees that could have logistic regression functions at the inner nodes and/or leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values. For this experimental setting of FT, we have used all the default parameters in WEKA workbench.

- REP: One of the simplest forms of pruning is reduced error pruning. Starting at the leaves, each node is replaced with its most popular class. If the prediction accuracy is not affected then the change is kept. While somewhat naive, reduced error pruning has the advantage of simplicity and speed. For this experimental set of REP, we have used all the default parameters in WEKA workbench.

- SMOTE: Regarding the use of the SMOTE pre-processing method [25], we consider only the 1-nearest neighbor (using the euclidean distance) to generate the synthetic samples and we balance both classes to the 50% distribution. For this experimental set of SMOTE, we have used all the default parameters in WEKA workbench.

- Logistic Regression (LR): Classifier for building linear logistic regression models. Logit Boost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection.

The results of the tenfold cross validation with standard deviation are shown in Table 2 to 5. The best two algorithms with the highest score for each dataset are highlighted in each row of every table. To better understand the relative performances of the examined algorithms we established a ranking schema of these algorithms based on the results of the pairwise comparisons.

More precisely, if an algorithm is significantly better than another it is credited with 1 point; if there is no significant difference between two algorithms then they are credited with 0.5 points; if an algorithm is significantly worse than another it is credited with 0 point. Thus, in the case m algorithms are examined, an algorithm that is significantly better than all the others for a given dataset is assigned a score of $m - 1$. In the last row of table 3-9 gives the rank and the numbers in parentheses indicate the number of significance points that the algorithm scores in a given dataset;

VI. Results and Analysis

We experimented with 6 standard datasets from the UCI repository (Arrhythmia, Credit-g, Liver, Glass, Hepatitis and Ionosphere); these datasets are standard benchmarks used in the context of learning. The goal is to examine whether the S2D learning framework from

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Section 3 achieve better predictive performance than a number of existing metric learning algorithms.

We compared the above methods with the C4.5, CART, FT, REP, SMOTE and LR (Linear Regression) state-of-the-art metric learning algorithms. In all the experiments we estimate accuracy using 10-fold cross-validation and control for the statistical significance of observed differences using t-test (sig. level of 0.05).

Table 3 Summary of tenfold cross validation performance for Accuracy on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	LR	S2D-C4.5	S2D-CART	S2D-LR
Arrhythmia	65.65±5.86	71.31±5.60	69.10±6.23	66.91±5.61	65.66±5.74	**72.77±5.47**	67.20±7.05	69.17±6.95	**72.50±6.11**
Credit-g	71.25±3.17	73.43±3.99	68.91±4.46	72.18±3.31	**76.50±3.39**	**75.37±3.53**	73.16±4.39	72.85±4.57	73.53±4.59
Glass	67.63±9.32	**71.26±7.84**	62.18±10.0	65.50±9.61	69.21±8.39	65.20±8.68	**70.09±13.9**	66.14±15.23	68.29±14.56
Hepatitis	79.22±9.57	77.10±7.12	81.90±8.38	78.62±7.13	78.35±9.09	**84.27±8.31**	82.38±11.2	80.41±11.49	**83.03±12.14**
Ionosphere	89.74±4.38	88.87±4.84	**90.26±4.97**	89.49±4.58	90.29±4.73	88.09±5.04	89.78±5.32	**90.19±5.27**	87.09±5.50
Waveform	75.25±1.90	76.65±1.75	82.86±1.72	76.52±1.71	80.47±1.47	**86.96±1.58**	76.80±1.89	77.20±2.01	**86.69±1.52**
Rank/Points	**VI (3.08)**	**V (3.25)**	**VI (3.08)**	**VII (2.91)**	**II (4.66)**	**I (5.33)**	**III (4.50)**	**IV (4.16)**	**I (5.33)**

4 Summary of tenfold cross validation performance for Tree size on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	LR	S2D-C4.5	S2D-CART	S2D-LR
Arrhythmia	80.62±5.95	20.80±4.22	**9.70±1.40**	22.00±5.68	81.64±5.78	- - - -	60.66±6.14	**19.19+5.63**	- - - -
Credit-g	126.85±20.66	24.46±18.09	31.72±4.30	76.81±19.60	152.41±24.36	- - - -	**20.48±4.58**	**13.28+8.52**	- - - -
Glass	46.16±4.58	21.16±9.14	**7.92±3.98**	19.70±5.47	47.62±5.68	- - - -	22.69±2.75	**15.17+5.37**	- - - -
Hepatitis	17.66±4.75	6.04±7.37	**3.22±0.75**	5.64±4.48	21.20±4.23	- - - -	7.18±2.30	**5.89+3.29**	- - - -
Ionosphere	26.74±3.89	8.42± 5.48	**5.72±1.08**	8.76±3.28	32.74±4.24	- - - -	18.47± 4.17	**7.01+4.63**	- - - -
Waveform	59.94±24.39	**98.32±33.96**	**77.48±8.22**	167.24±19.64	655.02±23.34	- - - -	417.26±28.75	138.72+68.33	- - - -
Rank/Points	**VI (1.0)**	**III (3.75)**	**I (5.5)**	**IV (3.33)**	**VII (0.0)**	**- - - -**	**V (2.5)**	**II (4.91)**	**- - - -**

Table 5 Summary of tenfold cross validation performance for AUC on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	LR	S2D-C4.5	S2D-CART	S2D-LR
Arrhythmia	**0.962±0.021**	0.817±0.063	0.785±0.069	0.786±0.068	0.780±0.065	0.896±0.050	0.802±0.08	0.807+0.068	**0.917±0.046**
Credit-g	**0.806±0.044**	0.716±0.055	0.65±0.075	0.705±0.054	0.778±0.041	**0.785±0.043**	0.68±0.067	0.687+0.113	0.741±0.064
Glass	0.409±0.272	**0.841±0.085**	0.784±0.101	0.827±0.080	0.808±0.101	0.831±0.077	0.776±0.18	0.818+0.158	**0.847±0.348**
Hepatitis	0.716±0.105	0.563±0.126	0.757±0.195	0.619±0.149	0.798±0.112	**0.852±0.142**	0.64±0.176	0.584+0.159	**0.823±0.199**
Ionosphere	0.850±0.066	0.896±0.059	0.900±0.060	0.902±0.054	0.904±0.053	**0.919±0.055**	0.892±0.07	0.901+0.058	**0.920±0.066**
Waveform	**0.972±0.021**	0.878±0.017	0.892±0.029	0.878±0.016	0.816±0.025	0.963±0.007	0.847±0.03	0.884+0.020	**0.965±0.008**
Rank/Points	**III (4.66)**	**IV (3.50)**	**VIII (2.91)**	**VII (3.00)**	**V (3.51)**	**II (6.75)**	**IX (1.83)**	**VI (3.08)**	**I (7.00)**

Table 6 Summary of tenfold cross validation performance for Error rate on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	LR	S2D-C4.5	S2D-CART	S2D-LR
Arrhythmia	34.34±5.86	28.69±5.60	30.90±6.23	33.09±5.07	34.34±5.74	**27.23±5.47**	32.80±7.05	30.83+7.75	**27.50±6.11**
Credit-g	28.75±3.17	26.57±3.99	31.09±4.46	27.82±3.31	**23.50±3.39**	**24.63±3.53**	26.84±4.39	27.16+4.75	26.47±4.59
Glass	32.37±9.31	**28.74±7.84**	37.82±10.0	34.50±9.61	30.79±8.39	34.81±8.68	**29.91±13.95**	33.86+15.82	31.71±14.56
Hepatitis	20.78±9.57	22.90±7.12	18.10±8.38	21.38±7.13	21.65±9.09	**15.73±8.30**	17.62±11.26	9.58+11.32	**17.02±12.13**
Ionosphere	10.26±4.38	11.14±4.84	9.74±4.97	10.51±4.58	**9.71±4.73**	11.91±5.04	10.22±5.32	**9.81+5.27**	12.91±5.50
Waveform	24.75±1.90	23.35±1.75	17.14±1.72	23.48±1.71	19.53±1.47	**13.04±1.58**	23.20±1.98	22.78+2.01	**13.31±1.52**
Rank/Points	**VI (3.08)**	**V (3.25)**	**VI (3.08)**	**VII (2.91)**	**II (4.66)**	**I (5.33)**	**III (4.50)**	**IV (4.16)**	**I (5.33)**

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

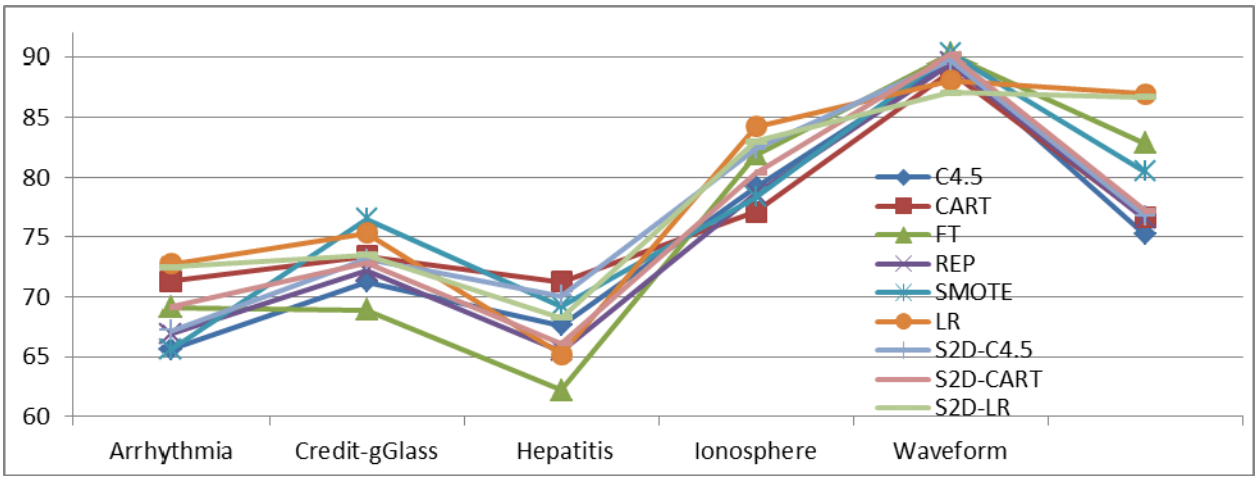


Fig.1. Trends in Accuracy for C4.5, CART, FT, REP, SMOTE, LR verses S2D-C4.5, S2D-CART and S2D-LR on UCI datasets

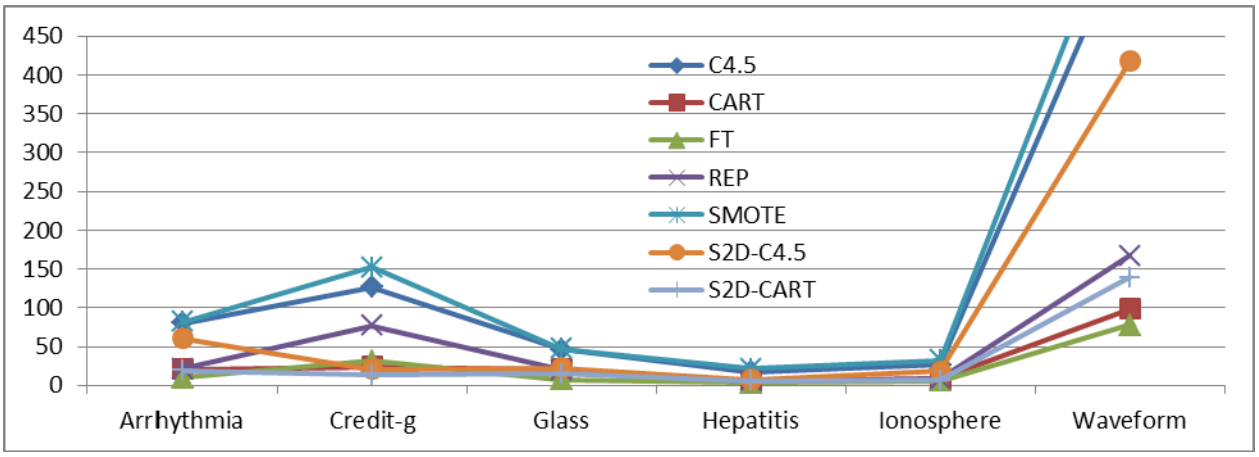


Fig.2. Trends in Tree Size for C4.5, CART, FT, REP, SMOTE verses S2D-C4.5 and S2D-CART on UCI datasets

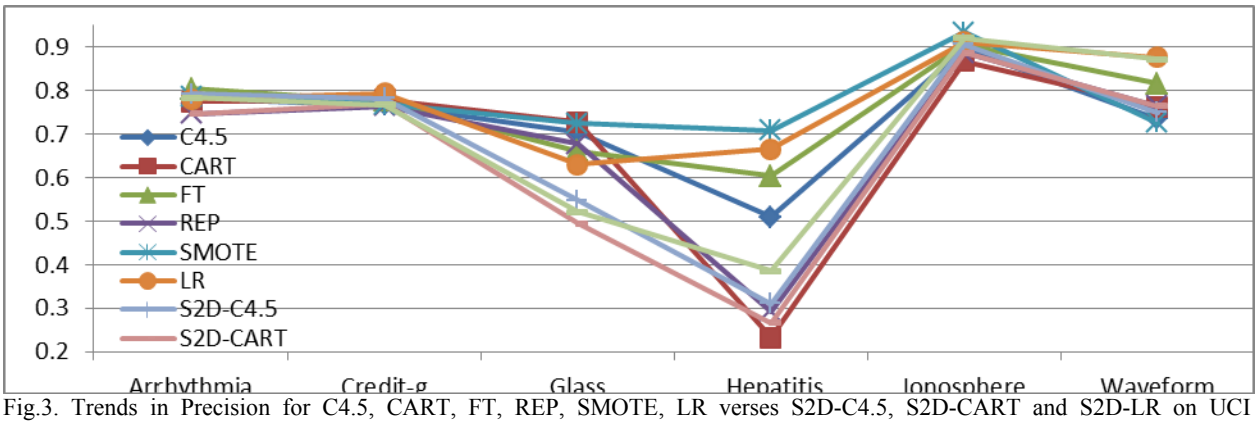


Fig.3. Trends in Precision for C4.5, CART, FT, REP, SMOTE, LR verses S2D-C4.5, S2D-CART and S2D-LR on UCI datasets

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

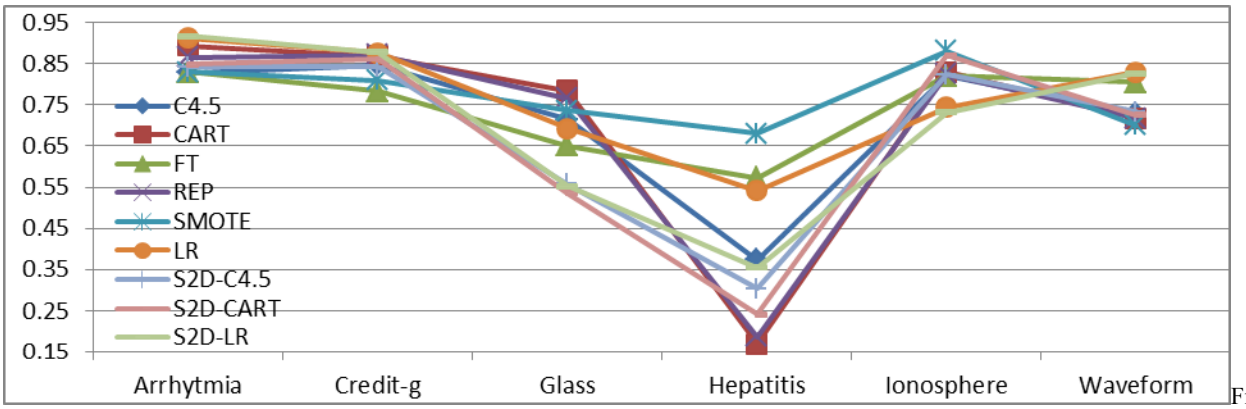


Fig.4. Trends in Recall for C4.5, CART, FT, REP, SMOTE, LR verses S2D-C4.5, S2D-CART and S2D-LR on UCI datasets

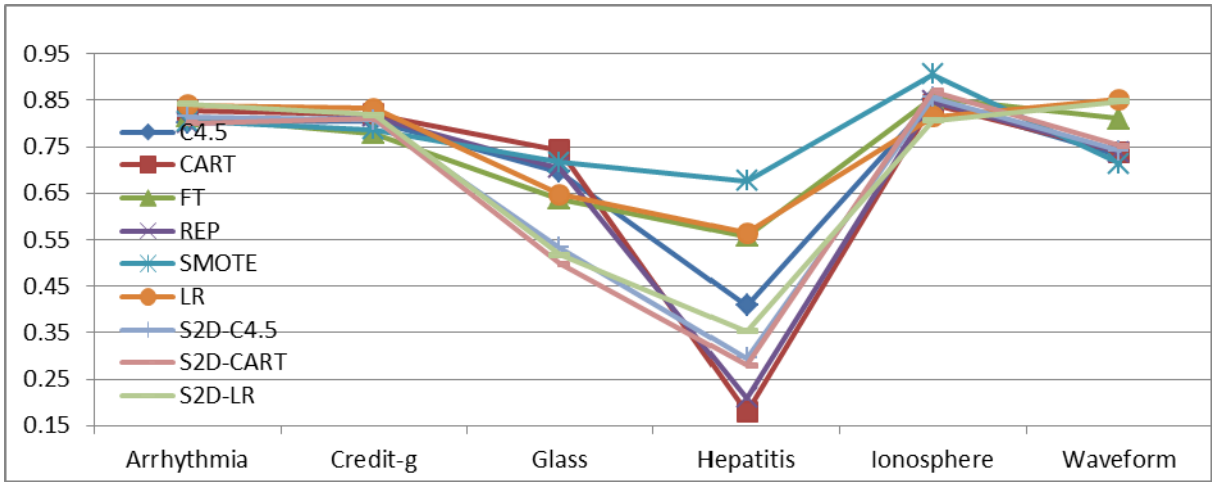


Fig.5. Trends in F-measure for C4.5, CART, FT, REP, SMOTE, LR verses S2D-C4.5, S2D-CART and S2D-LR on UCI datasets

Table 7 Summary of tenfold cross validation performance for Precision on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	LR	S2D-C4.5	S2D-CART	S2D-LR
Arrhythmia	0.781±0.065	0.774±0.06	**0.804±0.063**	0.747±0.066	0.786±0.059	0.781±0.052	**0.793±0.075**	0.747±0.074	0.783±0.069
Credit-g	0.767±0.025	0.779±0.030	0.776±0.033	0.765±0.025	0.768±0.034	**0.794±0.025**	**0.782±0.053**	0.769±0.038	0.767±0.037
Glass	0.705±0.141	**0.728±0.135**	0.661±0.141	0.677±0.138	**0.725±0.133**	0.630±0.132	0.547±0.262	0.495±0.171	0.521±0.269
Hepatitis	0.510±0.371	0.232±0.334	0.604±0.271	0.293±0.386	**0.709±0.165**	**0.666±0.307**	0.311±0.190	0.267±0.191	0.386±0.331
Ionosphere	0.895±0.845	0.868±0.096	0.906±0.080	0.886±0.092	**0.934±0.049**	0.911±0.076	0.906±0.086	0.887±0.097	**0.920±0.091**
Waveform	0.735±0.033	0.763±0.029	0.816±0.029	0.765±0.030	0.726±0.030	**0.876±0.025**	0.750±0.035	0.764±0.035	**0.872±0.028**
Rank/Points	**VI (3.50)**	**VI (3.50)**	**II (5.58)**	**VII (2.58)**	**III (4.83)**	**I (6.08)**	**IV (4.25)**	**VIII (1.41)**	**V (4.16)**

Table 8 Summary of tenfold cross validation performance for Recall measure on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	LR	S2D-C4.5	S2D-CART	S2D-LR
Arrhythmia	0.828±0.080	0.894±0.057	0.830±0.084	0.866±0.071	0.830±0.071	**0.913±0.065**	0.843±0.087	0.848±0.076	**0.917±0.064**
Credit-g	0.847±0.036	0.869±0.047	0.783±0.052	0.872±0.057	0.810±0.058	**0.876±0.039**	0.842±0.066	0.861±0.062	**0.879±0.051**
Glass	0.714±0.173	**0.784±0.176**	0.651±0.171	**0.766±0.166**	0.736±0.179	0.693±0.172	0.557±0.296	0.535±0.211	0.553±0.301
Hepatitis	0.374±0.256	0.169±0.236	**0.573±0.248**	0.187±0.239	**0.681±0.188**	0.542±0.276	0.304±0.222	0.241±0.175	0.355±0.344
Ionosphere	0.821±0.107	0.830±0.112	0.820±0.114	0.826±0.104	**0.881±0.071**	0.744±0.118	0.823±0.119	**0.871±0.096**	0.732±0.135
Waveform	0.728±0.036	0.717±0.037	0.805±0.030	0.717±0.043	0.701±0.038	**0.830±0.029**	0.733±0.042	0.725±0.039	**0.827±0.035**
Rank/Points	**VI (3.33)**	**III (4.41)**	**VII (3.08)**	**VI (4.25)**	**V (4.08)**	**I (5.58)**	**VII (3.08)**	**IV (3.50)**	**II (4.66)**

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 9 Summary of tenfold cross validation performance for F-measure on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	LR	S2D-C4.5	S2D-CART	S2D-LR
Arrhythmia	0.802±0.060	0.828±0.046	0.814±0.055	0.800±0.051	0.805±0.049	**0.840±0.046**	0.814±0.062	0.801±0.061	**0.842±0.045**
Credit-g	0.805±0.022	**0.820±0.028**	0.779±0.034	0.814±0.026	0.787±0.034	**0.833±0.025**	0.807±0.150	0.810±0.033	0.818±0.031
Glass	0.696±0.126	**0.743±0.124**	0.638±0.123	0.704±0.110	**0.716±0.125**	0.647±0.122	0.533±0.248	0.498±0.160	0.518±0.252
Hepatitis	0.409±0.272	0.179±0.235	0.557±0.207	0.210±0.259	**0.677±0.138**	**0.565±0.244**	0.295±0.187	0.280±0.162	0.353±0.314
Ionosphere	0.850±0.066	0.841±0.070	0.855±0.079	0.848±0.067	**0.905±0.048**	0.814±0.085	0.856±0.079	**0.870±0.070**	0.806±0.094
Waveform	0.730±0.025	0.738±0.023	0.810±0.021	0.739±0.025	0.713±0.028	**0.852±0.020**	0.741±0.030	0.753±0.028	**0.848±0.023**
Rank/Points	**VIII (3.00)**	**IV (4.16)**	**V (4.08)**	**VIII (3.00)**	**II (4.51)**	**I (5.83)**	**VI (3.75)**	**VII (3.16)**	**III (4.33)**

In Table 3-9, we present the results (with the score ranks) of the comparison between C4.5, CART, FT, REP, SMOTE, LR, S2D-C4.5, S2D-CART and S2D-LR (recall that the maximum score for an algorithm in a given dataset is 8). From these results we can make several observations. The developed S2D methods based on C4.5, CART and LR generally outperforms both FT, REP, SMOTE and traditional benchmarks; the advantage of our methods is most visible in the glass, hepatitis and ionosphere datasets. Finally, the methods that most often win are S2D-C4.5, S2D-CART and S2D-LR.

To quantify the relative performances of the different algorithms we computed for each method its average rank over all the examined datasets. These ranks are presented in the last row of Tables 3-9. From accuracy results in table 3, we observe that the best performance of 5.33 and better performances of 4.50 and 4.16 points are obtained for our variations S2D-LR, S2D-C4.5 and S2D-CART methods. From tree size results in table 4, one can observe that the second best performance of 4.91 points is obtained for the second variation of our S2D frame work. From AUC results in table 5, we observe that the best performance of 7.00 and better performances of 3.08 are obtained for our variations S2D-LR and S2D-CART methods. From error rate results in table 6, it is clearly noticeable that the best and second best performance is achieved by both of our algorithmic variations of S2D frame work.

From precision results in table 7, we observe that the better performance of 4.25 is obtained for our variation S2D-CART method. From recall results in table 8, one can observe that the second best performance of 4.66 points is obtained for our variation S2D-LR frame work. From F-measure results in table 9, we observe that the better performance of 4.33 is obtained for our variation S2D-LR method. Beyond this our framework of S2D has stood as best or second best performing algorithm in almost each and every datset.

These results are remarkable since S2D, which is based on a simple idea, performs equally well as the more elaborate metric learning algorithm that has been reported to consistently outperform other metric learning techniques over a number of non-trivial learning problems. Finally, we mention that the surprisingly little unexpected performance of S2D in terms of AUC in table 4, might be explained by the fact that its learning function is not convex and hence it is sensitive to the true positives or true negatives and false positives or false negatives.

Fig. 1 to 5 displays the accuracy, tree size, precision, recall and f-measure values for different UCI datasets for each learner verses S2D frame work. The y-axis in fig. 1-5 represents the accuracy, tree size, precision, recall and f-measure values respectively, with the x-axis representing the UCI datasets. Note that each figure is presented relative to a different scale to make the trends more visible, so care must be taken when comparing trends for different learners.

From, Fig. 1, we can observe that S2D-C4.5, S2D-CART and S2D-LR are the best algorithm, because on an average they had obtained the highest value in the ranking and the trend line representing these three algorithms can be seen to performing well on the chart. In order to analyze the tree size results, Fig. 2 shows the tree size on y-axis and UCI datasets on x-axis, where one can observe S2D variations had obtained the lowest tree size values and therefore good in the ranking and the trend line representing these three algorithms can be seen to performing well on the chart.

To analyze the precision, recall and f-measure results, Fig. 3-5 shows the precision, recall and f-measure on y-axis and UCI datasets on x-axis, where one can observe S2D variations had obtained the highest values and therefore good in the ranking and the trend line representing these three algorithms can be seen to performing well on the chart.

These results suggest that in the majority of the high dimensional datasets, the feature interactions are not important, and hence the methods that do not account for feature interactions have in general better performances. Alternatively, it might suggest that stronger regularization is needed. Moreover, it is interesting to note that the cases for which the good performance are difficult classification problems from the UCI datasets. This hints that there might be a bias of method development towards methods that perform well on UCI datasets; however, one can argue that they are really representative of the real world.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Scalability and computational complexity Issues:

Our approach S2D is more scalable in nature as one of the objective of our approach to handle highly scalable datasets by using feature selection approach. In our experiments we have tried for a dataset of size 150 to 5000.

The time complexity of our approach S2D can be obtained by combing the complexities of CFS and numbers of models build.

The time complexity of CFS is quite low. It requires $m((n2 - n)/2)$ operations for computing the pairwise feature correlation matrix, where m is the number of instances and n is the initial number of features. The feature selection search requires $(n2 - n)/2$ operations (worst case) for a forward selection or backward elimination search. Best first search in its pure form is exhaustive, but the use of a stopping criterion makes the probability of exploring the entire search space small. The total complexity for S2D is given as $m(n2\text{-}n)/2$ and S. Where, S is the total complexity of number of models build.

In overall, from all the tables we can conclude that our proposed method S2D have given good results when compared to benchmark algorithms. The unique properties of datasets such as size of the dataset and the number of attributes will also effect on the results of our S2D method. The above given results are enough to project the validity of our approach and more deep analysis should be done for further analysis.

VII. CONCLUSION

In this work, the emphasis is on the formalization of an efficient method for efficient learning of varied datasets. In terms of efficient learning methods we proposed similar-to-different (S2D) strategy and implemented the proposed approach on the top of three benchmark algorithms (C4.5, CART and LR). Furthermore, a comprehensive comparison study was conducted against seven benchmark metrics. The learning objective was to enhance learning capabilities among varied datasets. The study of the S2D variants on metric learning has confirmed that the learning objective was successfully accomplished. In terms of imbalance and balance metrics, very encouraging results have been achieved.

However, we believe that the unique properties of some datasets have negatively affected the S2D-based learning performances. We expect that using function approximation techniques to represent the learned policy will improve the S2D-based learning. This is the main purpose of our ongoing research work. In our future work, we will apply our proposed method for learning wide range of tasks, especially for high dimensional feature learning tasks.

REFERENCES

1. Ali Mirza Mahmood, Mohammad Imran, Naganjaneyulu Satuluri, Mrithyumjaya Rao Kuppa and Vemulakonda Rajesh., An Improved CART Decision Tree for Datasets with Irrelevant Feature, B.K. Panigrahi et al. (Eds.): SEMCCO 2011, Part I, LNCS 7076, pp. 539–549, 2011. Springer-Verlag Berlin Heidelberg 2011.
2. Ali Mirza Mahmood**,** Mrithyumjaya Rao Kuppa., EARLY DETECTION OF CLINICAL PARAMETERS IN HEART DISEASE BY IMPROVED DECISION TREE ALGORITHM, Second Vaagdevi International Conference on Information Technology for Real World Problems, 2010, IEEE Computer Society, DOI 10.1109/VCON.2010.12.
3. Ali Mirza Mahmood, Mrithyumjaya Rao Kuppa., A novel pruning approach using expert knowledge for data-specific pruning, Engineering with Computers (2012) 28:21–30, DOI 10.1007/s00366-011-0214-1.
4. Serkan ÇELİK, Veli TOPTAŞ. Vocabulary learning strategy use of Turkish EFL learners, Procedia Social and Behavioral Sciences 3 (2010) 62–71.
5. Zhou GuoDong, Zhang Min, Ji DongHong, Zhu QiaoMing. Hierarchical learning strategy in semantic relation extraction, Information Processing and Management 44 (2008) 1008–1021.
6. Edwin Lughofer. Hybrid active learning for reducing the annotation effort of operators in classification systems, Pattern Recognition 45 (2012) 884–896.
7. Kevin Duh, Akinori Fujino. Flexible sample selection strategies for transfer learning in ranking, Information Processing and Management 48 (2012) 502–512.
8. Xiaodong Yu , DexianHuang, YonghengJiang, YihuiJin. Iterative learning belief rule-base inference methodology using evidential reasoning for delayed coking unit, Control Engineering Practice 20 (2012) 1005–1015.
9. R.J. Gil, M.J. Martin-Bautista. A novel integrated knowledge support system based on ontology learning: Model specification and a case study, Knowledge-Based Systems 36 (2012) 340–352.
10. Md Nasir, Swagatam Das, Dipankar Maity, Soumyadip Sengupta, Udit Halder, P.N. Suganthan. A dynamic neighborhood learning based particle swarm optimizer for global numerical optimization, Information Sciences 209 (2012) 16–36.
11. Biao Niu, JianCheng, XiaoBai, HanqingLu . Asymmetric propagation based batch mode active learning for image retrieval, Signal Processing , Article in Press.
12. Ching-Hung Lee, Yu-Chia Lee. Nonlinear systems design by a novel fuzzy neural system via hybridization of electromagnetism-like mechanism and particle swarm optimisation algorithms, Information Sciences 186 (2012) 59–72.
13. Gwénolé Quellec, Mathieu Lamard, Michael D. Abràmoff, Etienne Decencière, Bruno Lay, Ali Erginay, Béatrice Cochener, Guy Cazuguel. A multiple-instance learning framework for diabetic retinopathy screening, Medical Image Analysis 16 (2012) 1228–1240.
14. Grana, C., Montangero, M., Borghesani, D., Optimal Decision Trees for Local Image Processing Algorithms, Pattern Recognition Letters (2012), doi: http://dx.doi.org/10.1016/j.patrec.2012.08.015.
15. Joel E. Denny, Brian A. Malloy, "The IELR(1) algorithm for generating minimal LR(1) parser tables for non-LR(1) grammars with conflict resolution", Science of Computer Programming 75 (2010) 943_979.
16. Sanjay Kumar Shukla a, M.K. Tiwari," Soft decision trees: A genetically optimized cluster oriented approach", Expert Systems with Applications 36 (2009) 551–563.
17. Sanjay Jain a,1, Steffen Lange b, Sandra Zilles, "Some natural conditions on incremental learning", Information and Computation 205 (2007) 1671–1684.
18. C. Grana, D. Borghesani, R. Cucchiara, Optimized Block-based Connected Components Labeling with Decision Trees, IEEE T Image Process 19 (2010) 1596–1609.
19. C. Grana, M. Montangero, D. Borghesani, R. Cucchiara, Optimal decision trees generation from or-decision tables, in: Image Analysis and Processing - ICIAP 2011, volume 6978, Ravenna, Italy, pp. 443–452.

20. H. Schumacher, K. C. Sevcik, The Synthetic Approach to Decision TableConversion, Commun ACM 19 (1976) 343–351.
21. Quinlan JR, C4.5: Programs for machine learning. Morgan,1993, Kaufmann, San Francisco.
22. Hall, M.A.: Correlation-based Feature Subset Selection for Machine Learning. PhD Thesis, Hamilton (1998)
23. A. Asuncion D. Newman. (2007). *UCI Repository of Machine Learning Database* (School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/ ~mlearn/MLRepository.htm
24. Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.
25. Chawla,N., Bowyer,K.,Kegelmeyer, P.:SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002).
26. Breiman L, Friedman J, Olshen R, Stone C (1984) Classification and regression trees. Wadsworth, Belmont.
27. Jesse Davis and Mark Goadrich (2006) "The Relationship between Precision-Recall and ROC Curves", in Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006.
28. Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48(1), 71–99.
29. Newport, E. L. "Constraints on learning and their role in language acquisition: Studies of the acquisition of american sign language". Language Sciences, (1988), 10(1), 147–172.
30. Newport, E. L. "Maturational constraints on language learning". Cognitive Science, (1990), 14(1), 11–28.
31. Dorigo, M., & Colombetti, M. "Robot shaping: An experiment in behavior engineering". (1998), MIT Press/Bradford Books.
32. Singh, S. P. "Transfer of learning by composing solutions of elemental sequential tasks". Machine Learning, (1992), 8, 323–339.
33. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**(4), 427–437 (2009),
34. Lemnaru, C., Potolea, R.: Imbalanced classification problems: systematicstudy, issues and best practices. In: Enterprise Information Systems, pp. 35–50. Springer, Berlin (2012)

AUTHORS PROFILE

Bhanu Prakash Battula received Master's Engineering degree on Computer Sciences & Technology in 2008 from Acharya Nagarjuna University and also received another Master's degree on Computer graduation. He is working as a Asst.Professor in the Department of Information Technology at Vignan's Nirula Institute of Technology and Science, Guntur,Andhra Pradesh. He published papers for International Journals. His research interests include Data Mining, Computer Security, Steganalysis and image processing.

R.Satya Prasad received PhD from Acharya Nagarjuna University in 2007.He is working as a Professor at Department of Science and Engineering, Acharya Nagarjuna University. He Published more than 23 National and International Publications and his research interests include Data Mining, Computer Security, Software reliability and Image Processing.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IMPLEMENTATION OF BACK PROPAGATION ALGORITHM FOR ESTIMATION OF STRESS AND STRAIN OF ALLOY WHEEL

[1]R.I.Rajidap Neshtar and [2]S.Purushothaman

[1]R.I.Rajidap Neshtar, Research Scholar
Department of Mechanical Engineering,
Vinayaka Missions University,
Salem, India.

[2]Dr.S.Purushothaman, Professor,
Department of Mechanical Engineering,
PET Engineering College,
Tirunelveli District-627117, India

***Abstract*—This paper presents estimation of stress and strain of a Rapid prototype product using artificial neural network (ANN). Back propagation algorithm is used to train the ANN topology. 3D model of alloy wheel is developed by using PROE. The model is analyzed using ANSYS to find the Von Mises stress and equivalent strain. The algorithm is trained using 15 values in the input layer of the ANN topology and two values in the output layer: stress and strain that are to be estimated during the testing stage of BPA algorithm. The number of nodes in the hidden layer for BPA varies depending upon the weight updating equations.**

***Keywords- Back Propagation Algorithm, Finite Element Method, Structural Analysis, Alloy Wheel, Mean Squared Error*.**

I. INTRODUCTION

Rapid prototyping plays an important role in manufacturing sample products for quick approval of the customers. Colors can be added to the products for aesthetic appearance to impress the customers. Different functionalities can be provided in the RP product and put into use to verify the usability. RP products are manufactured with layered Manufacturing. It is the process of deposition of material layer by layer. Continuous solid wire is melted, deposited on a non-sticky platen. The melt solidifies. During this process, the platen is moved in x, y directions as per a sequence and a sheet of solidified layered material is formed with a certain thickness. This is called first layer of deposition and the platen is lowered to receive the next layer of deposition. The process is repeated until, the entire height of the RP product is achieved. The material is not deposited during layering whenever holes, curves, slopes are to be provided in the product.

The sequence of RP manufacturing is presented as follows:

Step 1:	Created computer aided drafting and design of the RP product.
Step 2:	Analyse the model using Finite element software like ANSYS, NASTRAN, if the RP product is the end product. Otherwise, if the product is only for a display or customer approval process, no analysis is required for small sized RP product.
Step 3:	Transfer the final RP model to RP machine. The software in the RP machine will convert the model suitably for layered manufacturing.

In step 2, an alternative procedure is proposed. After analyzing the RP product with FEM software, the outputs of the FEM software along with the inputs of FEM are used as training patterns for the proposed artificial neural network algorithms. The algorithms learn the input output relationship of the values used an input and obtained as outputs for analyzing the RP product. This type of analysis can be extended to all types of RP products as well as other engineering products.

1 Companies working on unique products and companies which modify the size of a particular RP product can avail the proposed.

2. Such companies need not obtain costly FEM software; instead they can use the proposed approach to save the investment of the FEM software for obtaining the stress, strain values of the RP product.

II. REVIEW OF LITERATURE

Wenbin et al, 2002, applied optimization algorithms like genetic algorithm for the outputs of FEM. Attention is devoted to examining the effects of critical geometric features on the stress distribution at the interface between the blade and disk using a feature-based geometry modeling tool and the optimization techniques. Various aspects of this problem are presented: (1) geometry representation using ICAD and transfer of the geometry to a finite element analysis code, (2) application of boundary conditions/loads and retrieval of analysis results, (3) exploration of various optimization methods and strategies including gradient-based and modern stochastic methods. A product model from Rolls-Royce is used as a base design in the optimization.

Manevitz et al, 2005, Basic learning algorithms and the neural network model are applied to the problem of mesh

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

adaptation for the finite-element method for solving time-dependent partial differential equations. Time series prediction via the neural network methodology is used to predict the areas of ''interest'' in order to obtain an effective mesh refinement at the appropriate times. This allows for increased numerical accuracy with the same computational resources as compared with more ''traditional'' methods.

Toraño et al, 2008, used neural networks to estimate stress –strain of long mining wall. This knowledge and the detailed structural and constructive characteristics of the support systems allow the simulation of the behavior of the roof supports through finite element method.

Mohsen Ostad Shabani et al, 2011, used ANN and FEM for estimating the yield stress, Ultimate Tensile Strength, maximum force and elongation percentage of solidification in A356 alloy. The prediction of ANN model with one output was found to be in good agreement with experimental data. The results show that the prediction of neural network modelling with four outputs cannot really give a good performance and show the best relationship between each individual output and its inputs.

Mohsen Ostad Shabani et al, 2012, developed ANN model to predict the hardness, yield stress, ultimate tensile strength and elongation percentage. The prediction of ANN model was found to be in good agreement with experimental data. It is concluded that considerable savings in terms of cost and time could be obtained by using neural network model.

Hyuntae et al, 2013, developed noncommercial ANN simulator with graphical user interface (GUI) to assist in rapid data modeling and analysis in the engineering diffraction field. This software has been used to train and simulate the finite element modeling (FEM) data for a fiber composite system, both forward and inverse. The forward neural network simulation precisely reduplicates FEM results several orders of magnitude faster than the slow original FEM. The simulator GUI also reveals that output node size for materials parameter and input normalization method for strain data are critical train conditions in inverse network..

III. METHODOLOGY

ANSYS 14 software is used for analyzing the rapid prototypes. An alloy wheel is considered for analyzing the estimation accuracy of artificial neural network (ANN) algorithm. The analysis conditions of the wheel is presented. Load deformation response characteristics of the wheel are determined using a finite element computer analysis. Results of this analysis is used as training patterns for the ANN algorithm.

Artificial neural network algorithm is used to supplement the estimation of stress and strain values of the RP models. The result of the analysis of alloy wheel is obtained both in graphical and in numerical values. The numerical values of stress and strain are used to train the artificial neural network (ANN) topology by using Back propagation algorithm (BPA).

The ANN is trained using 15 values in the input layer two values in the output layer: stress and strain that are to be estimated during the testing stage of ANN algorithms. The number of nodes in the hidden layer varies depending upon the weight updating equations. Exact number of nodes, is fixed based on the trial and error method, in which the accuracy of estimation by the BPA is used as the criteria for the performance of ANN algorithm.

The training of patterns used for the ANN are chosen from the stress and strain data generated using ANSYS software. During the training process, mesh node numbers are presented in the input layer of the ANN and correspondingly, stress and strain values are presented in the output layer of the ANN. Depending upon the type of values present in the patterns, the learning capability of the ANN algorithms varies.

A. Back Propagation Algorithm (BPA)

The concept of steepest-descent method is used in BPA to reach a global minimum. The number of layers are decided initially. The number of nodes in the hidden layers are decided.

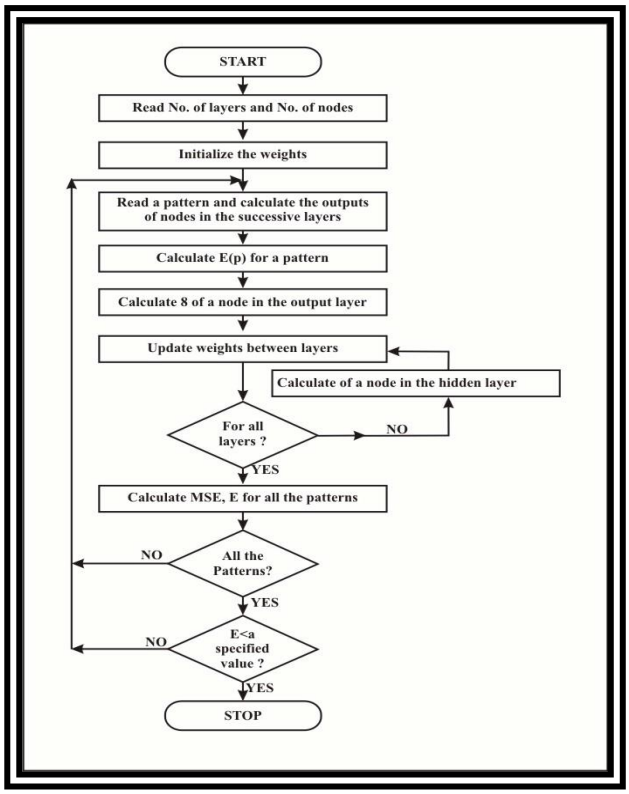


Figure 1 Flow-chart of BPA

It uses all the 3 layers (input, hidden and output). Input layer uses 15 nodes, hidden layer has 2 nodes and the output layer includes two nodes.

Random weights are used for the connections between nodes. Error at the output layer of the network is calculated by presenting a pattern to the input layer of the network. Weights are updated between the layers by propagating the error backwards till the input layer. All the training patterns are

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

presented to the network for learning. This forms one-iteration. At the end of iteration, test patterns are presented to ANN and the prediction performance of ANN is evaluated. Further training of ANN is continued till the desired prediction performance is reached.

The concept of steepest-descent method is used in BPA to reach a global minimum. The number of layers are decided initially. The number of nodes in the hidden layers are decided. It uses all the 3 layers (input, hidden and output). Flow-chart for BPA is shown in Figure 1.

B. Steps Involved In Training Bpa

Forward Propagation

The hidden layer connections of the network are initialized with weights.

The inputs and outputs of a pattern are presented to the network.

The output of each node in the successive layers is calculated by using equation (1).

$$O_{(\text{output of a node})} = 1/(1+\exp(-\sum w_{ij}x_i)) \quad (1)$$

For each pattern, error is calculated using equation (2).

$$E(p) = (1/2)\sum(d(p) - o(p))^2 \quad (2)$$

Reverse Propagation

For the nodes, the error in the output layer is calculated using equation (3).

$$\delta_{(\text{output layer})} = o(1-o)(d-o) \quad (3)$$

The weights between output layer and hidden layer are updated by using equation (4).

$$W_{(n+1)} = W_{(n)} + \eta\, \delta_{(\text{output layer})}\, o_{(\text{hidden layer})} \quad (4)$$

The error for the nodes in the hidden layer is calculated by using equation (5).

$$\delta_{(\text{hidden layer})} = o(1-o)\sum\delta_{(\text{output layer})}W_{(\text{updated weights between hidden \& output layer})} \quad (5)$$

The weights between hidden and input layer are updated by using equation (6).

$$W_{(n+1)} = W_{(n)} + \eta\, \delta_{(\text{hidden layer})}\, o_{(\text{input layer})} \quad (6)$$

The above steps complete one weight updation.

The above steps are followed for the second pattern for subsequent weight updation. When all the training patterns are presented, a cycle of iteration or epoch is completed. The errors of all the training patterns are calculated and displayed on the monitor as the MSE.

$$E_{(MSE)} = \sum E_{(p)} \quad (7)$$

IV. RESULTS AND DISCUSSIONS

Static analysis calculates the effects of steady loading conditions on a structure. Static analysis, however, includes steady inertia loads (such as gravity and rotational velocity), and time-varying loads that can be approximated as static equivalent loads. Static analysis is used to determine the displacements, stresses, strains, and forces in structures or components caused by loads that do not induce significant inertia and damping effects. Steady loading and response conditions are assumed; that is, the loads and the structure's response are assumed to vary slowly with respect to time.

Bounding Box	
Length X	0.2 m
Length Y	0.5 m
Length Z	0.5 m
Properties	
Volume	5.2943e-003 m³
Mass	41.561 kg
Centroid X	-8.688e-002 m
Centroid Y	4.5371e-005 m
Centroid Z	-3.5088e-005 m
Moment of Inertia Ip1	1.4413 kg·m²
Moment of Inertia Ip2	0.80226 kg·m²
Moment of Inertia Ip3	0.80168 kg·m²
Statistics	
Nodes	22941
Elements	12652

Structural Steel > Constants

Density	7850 kg m^-3
Coefficient of Thermal Expansion	1.2e-005 C^-1
Specific Heat	434 J kg^-1 C^-1
Thermal Conductivity	60.5 W m^-1 C^-1
Resistivity	1.7e-007 ohm m

Figure 2 shows the amount of von–Mises stress in (Pa) presented at various nodes mentioned in the x-axis. Figure 3 presents strain on the alloy wheel.

The node number is given as one of the inputs at the input layer of the ANN topology. Based on the requirements of the outputs, number of input parameters can be increased. Table 2 presents the numbers of nodes used in the input layer, hidden layer and output layer during training.

ANSYS 14 software is used for analyzing the alloy wheel. The numerical values of stress are used to train the artificial neural network (ANN) topology by using Back propagation algorithm (BPA). The training of patterns used for the ANN algorithms are chosen from the strain data generated using ANSYS program. During the training process, node numbers are presented in the input layer of the ANN and correspondingly, strain values are presented in the output layer of the ANN. Depending upon the type of values present in the patterns, the learning capability of the ANN algorithms varies.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Figure 7-9 present performance of BPA in estimating the stress and strain.

V. CONCLUSION

In this paper, ANSYS 14 software is used for analyzing the alloy RP model. The experimental simulation and the conditions under which the experiments were simulated are mentioned. Load deformation response characteristics of alloy wheel model is determined using a finite element computer analysis.

Artificial neural network algorithms have been used to supplement the estimation of stress and strain values of the proposed alloy wheel model. The result of the analysis of the model is obtained both in graphical and in numerical values. The numerical values of stress and strain are used to train the artificial neural network (ANN) topology by using Back propagation algorithm (BPA).

The algorithm is trained using 15 values in the input layer of the ANN topology and two values in the output layer: stress and strain that are to be estimated during the testing stage of BPA algorithm. The number of nodes in the hidden layer for BPA varies depending upon the weight updating equations.

1. As the number of training patterns increase, the time taken by BPA to converge also increases.
2. The optimum number of nodes in the hidden used for BPA is 2 nodes. As the number of nodes in the hidden layer increases, the accuracy of BPA for stress estimation and strain estimation reduces.

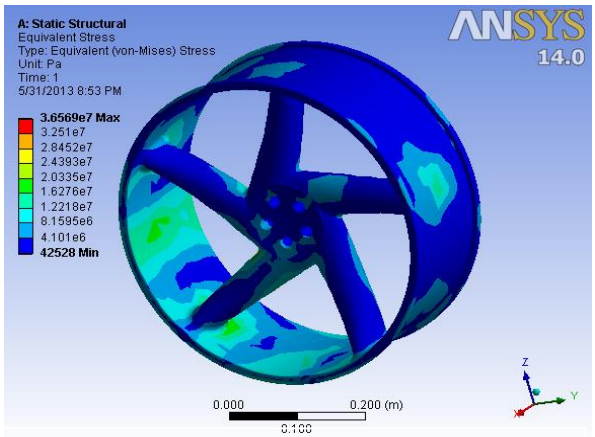


Figure 2 Alloy Wheel with stress distributed

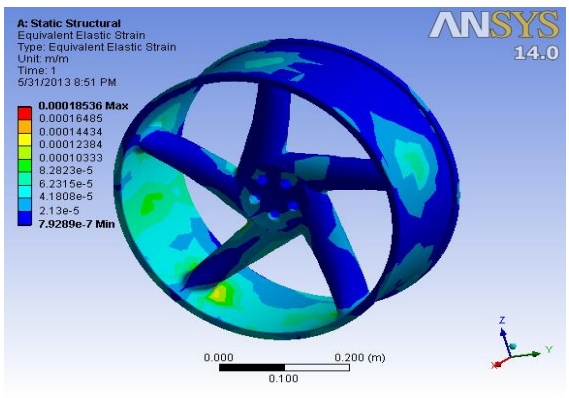


Figure 3 Equivalent strain

Figure 4 stress distribution on the form of plot. Figure 5 and Figure 6 present data used for training BPA.

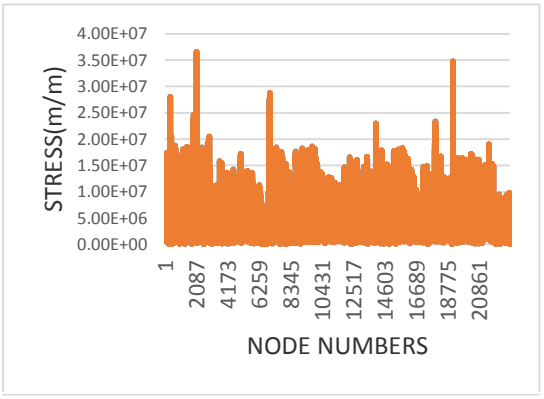


Figure 4 Stress distribution of alloy wheel

Table 1 Training parameters for the proposed ANN algorithms	
Inputs to the ANN topology	***Target outputs for the ANN topology***
1. Diameter / Length of the model 2. Thickness of the model 3. Width of the model 4. Number of holes 5. Number of ribs 6. Maximum to minimum diameter / Length 7. Uniformly distributed load (1='yes' , 0='no') 8. Uniformly distributed load value 9. Point load (1='yes' , 0='no') 10. Point load value 11. Yield strength 12. Ultimate strength 13. Young's modulus 14. Poisson ratio 15. Node number	1. Equivalent von-Mises stress in (Pa) 2. Equivalent strain (m/m)

Table 2 ANN topology			
	Nodes in the input layer	***Nodes in the Hidden layer***	***Nodes in the output layer***
BPA	15	3	2

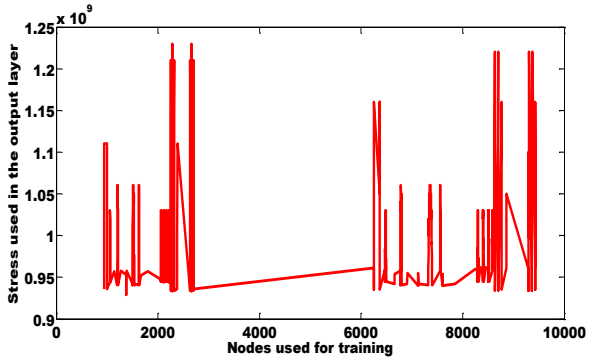


Figure 5 Stress data for training ANN

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

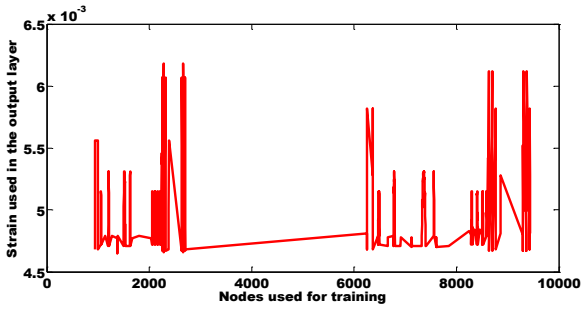


Figure 6 Strain data for training ANN

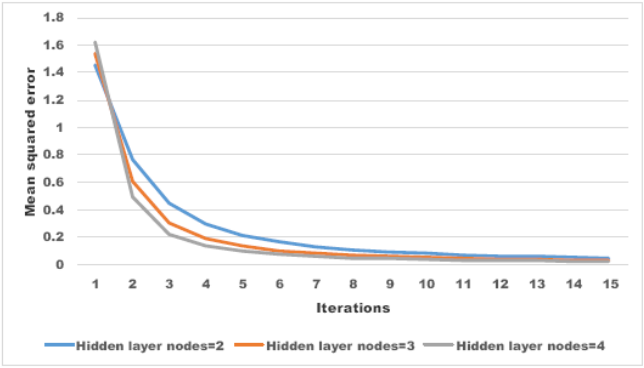


Figure 7 Convergence rate of BPA for learning strain

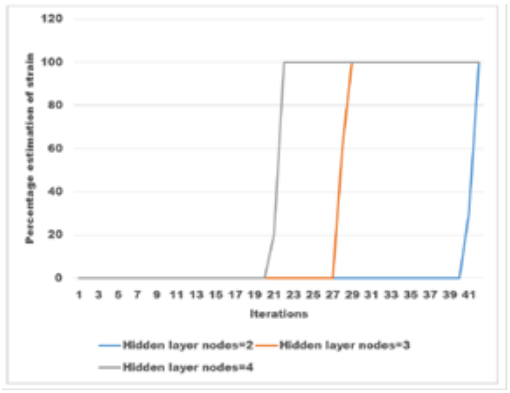


Figure 8 Percentage number of nodes–strain estimated

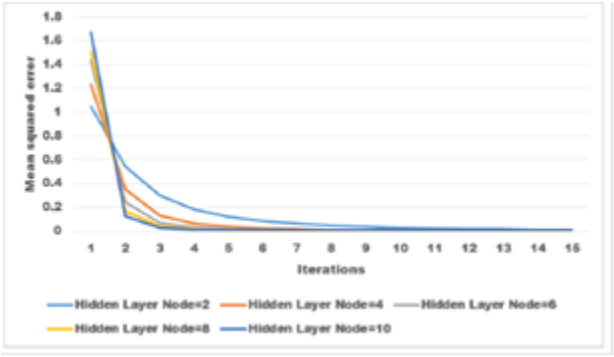


Figure 9 Convergence rate for BPA while training stress data

REFERENCES

[1] Mohsen Ostad Shabani, Ali Mazahery, Mohammad Reza Rahimipour and Mansour Razavi, 2012, FEM and ANN investigation of A356 composites reinforced with B4C particulates, Journal of King Saud University – Engineering Sciences, Vol.24, pp.107–113.

[2] Larry Manevitz, Akram Bitar and Dan Givoli, 2005, Neural network time series forecasting of finite-element mesh adaptation, Neurocomputing, Vol.63, pp.447–463.

[3] Javier Toraño, Isidro Diego, Mario Menéndez and Malcolm Gent, 2008, Finite element method (FEM) – Fuzzy logic (Soft Computing) – virtual reality model approach in a coalface long wall mining simulation, Automation in Construction, Vol.17, pp.413–424.

[4] Mohsen Ostad Shabani and Ali Mazahery, 2011, The ANN application in FEM modeling of mechanical properties of Al–Si alloy, Applied Mathematical Modelling, Vol.35, pp.5707–5713.

[5] Wenbin Song, Andy Keane, Janet Rees, Atul Bhaskar, and Steven Bagnall, 2002, Turbine blade fir-tree root design optimisation using intelligent CAD and finite element analysis, Computers and Structures, Vol.80, pp.1853–1867.

[6] Hyuntae Na, Seung-Yub Lee, Ersan Üstündag, Sarah L. Ross, Halil Ceylan, and Kasthurirangan Gopalakrishnan,Development of a Neural Network Simulator for Studying the Constitutive Behavior of Structural Composite Materials, ISRN Materials Science, Vol.2013, Article ID 147086, 10 pages.

AUTHORS PROFILE

R.I.Rajidap Neshtar completed his B.Tech in Mechanical Engineering and M.E in CAD/CAM from Karunya University, Coimbatore in 2008. He has 5 years of teaching experience. Presently he is working as an Assistant Professor in Bethlahem Institute of Engineering, India. He has attended various Faculty development programs and Soft skill trainings and attended various national and international conferences.

Dr.S.Purushothaman completed his PhD from Indian Institute of Technology Madras, India in 1995. He has 124 publications to his credit. He has 19 years of teaching experience. Presently he is working as Professor in PET college of Engineering, India

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

An Approach to Reveal Website Defacement

Rajiv Kumar Gurjwar
Computer Science Engineering Department
MANIT, Bhopal (M.P.), INDIA

Divya Rishi Sahu
Computer Science Engineering Department
MANIT, Bhopal (M.P.), INDIA

Deepak Singh Tomar
Computer Science Engineering Department
MANIT, Bhopal (M.P.), INDIA

***Abstract*— Due to ad-hoc nature of web application development and design complexity of web application, it is difficult to attain fool proof web security. In recent years invaders defaced several web sites by projecting techniques such as phishing, code injection etc. In the web defacement attack the invader changes the visual appearance of the webpage. The business competitor, Insurgent and extremist groups defame the reputation of the organizations and mislead public through these types of attacks. Manual monitoring and scrutinizing these attacks on web sites is a time consuming and tedious task for law enforcement agencies. Hence there is a need to develop a system which effectively monitors the content of web sites and automatically generate alarm for any suspicious or threatening activity. In this work a prototype system is developed to scrutinize and detects the defacement activities automatically. At first phase web contents are preprocessed and stored in the web domain dictionary. Second phase checked integrity of web contents through CRC32, MD5, SHA 512, PSNR and SSIM techniques. The developed system successfully scrutinizes the web defacement attacks and it would be helpful for web administrator to monitor the web defacement attacks.**

Keywords-web security; website defacement; internet security

I. INTRODUCTION

Websites are prone to be attacked by campaign hackers because of complexity in achieving full proof security during designing phase of a web site. The targets are mainly government websites, official agencies and trade groups. Attacks on websites cover malicious activities such as phishing [1], code injection [2], Website Defacement and many more. In which *website defacement* is an exploitation of the techniques to alter the content of web pages by suspicious user. In this the attacker defaces the reputation of an organization by modifying the content of home page. Recently several such attacks on websites are encountered in which text content and web images of reputed organizations are altered to spoil their reputation

Loop holes such as less protection, improper configuration of web pages, providing weak passwords and availing access of web page contents to all users with full privilege leads to web defacement attacks.

Web defacement can be broadly categorized into Text Defacement and Image Defacement. For the detection of text defacement ample of work has been done using numerous text integrity techniques, whereas web image defacement is a new research field. In this work main emphasis is given on detecting website defacement in context of web images using CRC 32 checksum, hashing, and PSNR & SSIM techniques [3, 4]. Experimental evaluation is done to determine accuracy of the proposed work.

This paper is organized as follows: Section 2 states recent website defacement incidences. In section 3 set of techniques that have been explored in detection of website defacement are reviewed. Section 4 describes detailed architecture of developed prototype system. Section 5 defines experimental setup including experimental dataset for result evaluation. Section 6 discusses the evaluation of methods.

II. RECENT WEBSITE DEFACEMENT

The researchers, security experts and law enforcement agencies had pointed out various web defacement cases:

Recently, Chinese websites has been defaced in Anonymous attack [5]. In this defacement the Anonymous hacking group claims to have defaced almost 500 websites in China. Targets hit in the mass defacement included government sites, its official agencies, trade groups and many others. A message put on the hacked sites said the attack was carried out to protest against the Chinese government's strict control of its citizens.

A group of purportedly Gazan hackers defaced Israel's Fire and Rescue Services website [6] on January 13, 2012. They added a "death to Israel" message to the website and modified a picture of Israel's Deputy Foreign Minister, and superimposing foot prints over his face.

Two Government website of Department of Transportation and Communication Philippines were defaced by Turkish Hackers [7], October 8, 2012. Hack was meant to protest rights violations.

The law enforcement agencies are trying their best to catch such cases. Manual monitoring and scrutinizing these attacks on website is still a hectic and time consuming task. Hence there is a need to develop a system which effectively monitors the content of web sites and automatically generate an alarm to reveal any suspicious or threatening activity.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. BACKGROUND AND LITERATURE SURVEY

This section gives a brief introduction to techniques that have been proposed in past years by researchers to develop a worthwhile detection system for website defacement.

Alberto Bartoli et al. [8] proposed a framework for Large-Scale Detection of Website Defacements which is based on Anamoli detection technique. In this technique the network during the training period to define normal traffic becomes unprotected from attack and to associate an alarm with the specific event that triggered the alarm is very difficult task. These are the limitations of this technique. The proposed prototype system overcomes these limitations. It generates the alarm for any suspicious activity rather than specific event.

Tushar Kanti et al. [9] enforced a website defacement detection mechanism at browser level. Pan Shi et al., Sobey et al. and Jackson et al. [10-12] confirmed that users generally do not notice Extended Validation (EV) certificates in Firefox and similarly, it is also found that the design in IE failed to protect users from phishing attack. Being attention grabbing is the first necessity of a security in web browsers.

An XML technology-based solution WALSG (Web Application Level Security Gateway) to web security is proposed in [14]. Their solution WALSG does not provide security to all web pages of website but proposed system in this paper does.

IV. PROPOSED FRAMEWORK

In this work a framework is proposed for automatically detection of web content tampering. The proposed solution first extracts web components and then generates the alert by checking the integrity of extracted components with original contents stored in web repository. The architecture for proposed framework is depicted in Fig. 1.

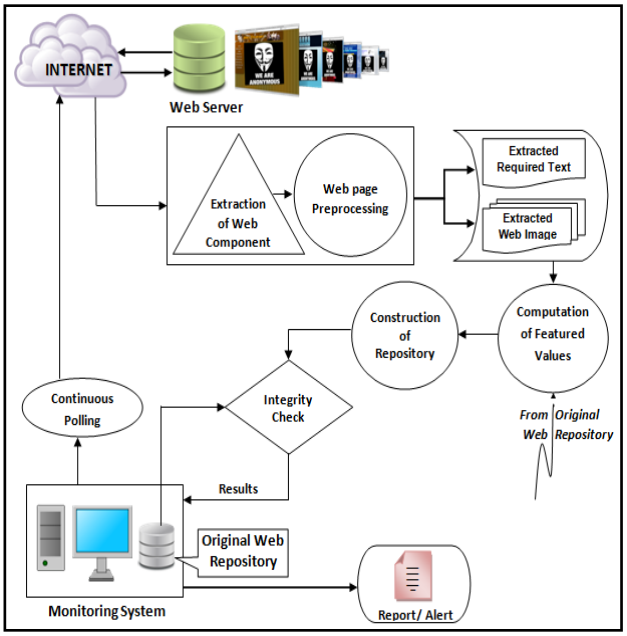


Figure 1. Architecture of Proposed Framework

This framework has five modules named as Monitoring System, Extraction of web page components, Pre-processing of Text and Images, Integrity Verification and Web Content Repository, which are illustrated as follows:

A. Monitoring System

Monitoring system is used by administrator of website to watch all the reports generated after polling mechanism. It contains original web repository of website. This web repository is regularly updated whenever any updates happen in website, either by administrator or by authentic user at regular period of time.

To make this monitoring system more effective a polling mechanism is required. Polling is a technique to trigger the website defacement detection mechanism at regular interval. Common polling techniques are: Bench Mark polls, Brushfire polls and Tracking polls [25]. Benchmark poll is generally the first poll taken in consideration. It lets us know the standard where we stand. It helps to know how to use the limited resources effectively in limited time. Brushfire polls are taken during the period between the Benchmark polls and Tracking polls. In tracking poll, polling is repeated periodically with the same group of data to detect changes happened or not. In this work tracking polling mechanism is explored.

B. Extraction of Web Components

To detect the defacement in text and image the first task is to extract web contents such as Front End Elements, images, Back End elements, downloadable files, CSS files, java script files, HTML headers along with text from a web site. The main task of this phase is to extract these web contend from web server.

C. Web page Pre-processing

The next step after web page extraction is to preprocess them. Web pages contains heterogeneous data such as text and multimedia contents in which only text and images are required for the detection techniques used, rest are irrelevant. Web page pre-processing [15] is the process to eliminate these irrelevant contents. In this module, text and image pre-processing are explored.

1) Text Pre-processing

Text Preprocessing aims to removes all the java script code and all the HTML tags from extracted web page.

```
<html>
<title> Abash | WordMemorize </title>
                <div><ul id="menu-top-menu" ><a>Blogs</a>
<ul><li ><a>Word Memorize</a></li>
    <li id="menu-item-180"><a>Books For GATE</a></li></ul></li>
<li id="menu-item-188"><a>Free_SMS</a></li></ul></div>
          <ul id="menu-main-menu"><li><a>What's New Today</a></li>
<li id="menu-item-209"><a>Words</a>
<ul><li><a>A-E</a></li>
    <li><a>F-J</a></li>
    <li><a>K-O</a></li>
    <li><a>P-T</a></li>
    <li><a>U-Z</a></li></ul></li>
<li><a>Exams</a></li>
<li><a>Downloads</a></li>
```

Figure 2. Input HTML page -bash_WordMemorize.htm

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The output of this phase is the actual content for further module.

This text is saved in text file having same name as of HTML page. A sample training web page is shown in Fig. 2 and its pre-processed text is shown in Fig. 3.

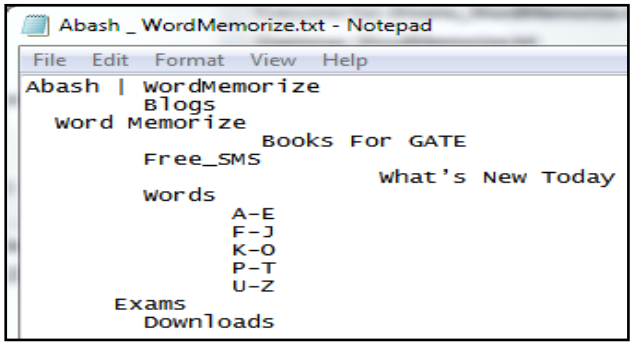


Figure 3. Extracted text of input page

2) Image Pre- processing

In this phase only web image is taken out from web page and rest multimedia components are eliminated. The web browser fetch the image using image path written in "src" attribute of image tag. Fig. 4 depicts the pre-processing of image.

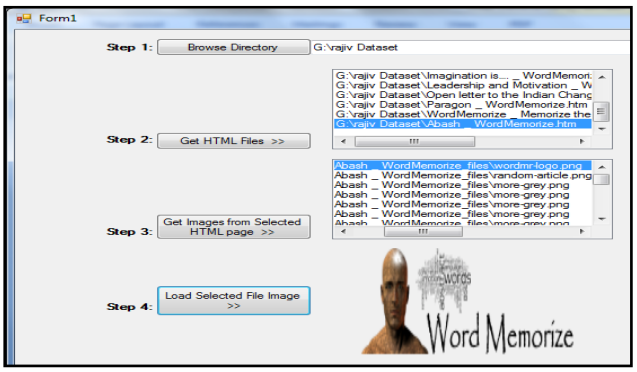


Figure 4. Extraction process of image from HTML page

D. Integrity Techniques

Integrity is the quality of being identical. In terms of website defacement, integrity is the quality of having same website contents (i.e. text as well as multimedia contents). In this module Text Integrity and Image Integrity are tested.

Text Integrity is the process to detect the alteration of text. It is done by comparing text character-by-character. It requires at least one file contains unaltered or original text.

Image Integrity is the process to detect any alteration in pixels of image. In this images are compared pixel by pixel. Two images are called same when each and every pixel has same value.

Techniques such as CRC 32, MD5, PSNR, and SSIM are earlier devised approaches to verify integrity of text and images. But they suffer from certain short comes like CRC 32 has security problem, MD5 has limited output message digest so it has collision problem. PSNR and SSIM requires large backup database to store original images for calculation and fail to check text integrity. To overcome these limitations SHA512 hashing method is explored here. Techniques used to check image integrity are discussed as follows:

1) Peak Signal-to-Noise Ratio (PSNR)

It is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise of image. It is used to compare results from the Images. [22-24]. In this framework, it is used to verify integrity using two web images in which one is original image and other may be altered image. PSNR is usually expressed in terms of the logarithmic decibel scale.

Equation (1) is used to calculate the value of PSNR which is given by:

$$\begin{aligned} \mathrm{PSNR} &= 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) \\ &= 20 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I}{\sqrt{\mathrm{MSE}}}\right) \\ &= 20 \cdot \log_{10}(\mathrm{MAX}_I) - 10 \cdot \log_{10}(\mathrm{MSE}) \end{aligned} \quad (1)$$

{Here,

MAX_I = maximum possible pixel value of the image

MSE = mean squared error }

When the pixels are represented using 8 bits per sample, then MAX_I is 255. To find MSE in (3) first mean of random sample of size n from a population, $X_1 \ldots X_n$ is calculated which is given in (2).

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i) \quad (2)$$

$$\mathrm{MSE}(\overline{X}) = \frac{\sigma^2}{n} \quad (3)$$

{Here,

σ^2 = Variance }

Equation (4) is used to calculate the Variance which is given by:

$$\text{Variance } (\sigma^2) = \frac{1}{n}\sum_{i=1}^{n}(i - (\overline{X})^2) \quad (4)$$

This method returns numeric values as illustrated by Table I. Infinity (∞) means image is unaltered.

TABLE III. PSNR VALUES

Original Image		Tampered Image	PSNR Value
Name	Size (in pixel)		
image1.png	400x300	Minute distortion	48.9329
image2.png	400x300	Heavy distortion	15.8742
image3.png	400x300	No Change	∞

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

2) Structural Similarity (SSIM)

SSIM is designed to improve on traditional methods like peak signal-to-noise ratio (PSNR) and mean squared error (MSE), which have proven to be inconsistent with human eye perception [20, 21]. The SSIM metric is calculated on various windows of an image. To measure $SSIM(x, y)$ between two windows x and y of common size $N \times N$ is given in (5):

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2+\mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (5)$$

{Here,

- μ_x = the average of x;
- μ_y = the average of y;
- σ_x^2 = the variance of x;
- σ_y^2 = the variance of y;
- σ_{xy} = the covariance of x and y;
- $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$ two variables to stabilize the division with weak denominator;
- L = dynamic range of the pixel-values (typically this is $2^{\#bits\ per\ pixel} - 1$) i.e. $L = 255$;
- $k_1 = 0.01$ and,
- $k_1 = 0.03$ }

Equation (5) is derived with μ_x & μ_y average or mean which can be calculated in similar manner like in (2), σ_x^2 & σ_y^2 is calculated like (4) and σ_{xy} can be calculated with (6):

$$Covariance\ (\sigma_{xy}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad (6)$$

{Here,

n = sample size

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y }

This method returns numeric values as illustrated by Table II. In this SSIM values 1 means image is unaltered.

TABLE IV. SSIM VALUES

Original Image		Tampered Image	SSIM Value
Name	**Size (in pixel)**		
image1.png	400x300	Minute distortion	0.9985
image2.png	400x300	Heavy distortion	0.8347
image3.png	400x300	No Change	1

Both implemented methods (i.e. PSNR and SSIM) are not able to check text integrity. Both are used for image integrity verification and image assessment [18].

3) Integrity through CRC32 Checksums

In this approach CRC checksum [13] is calculated for text file as well as image files verification. Sample CRC32 checksums are shown by:

file_one.txt	c45ad668
file_two.txt	7903b8e6
logo1.png	e99a65fb

CRC32 is not suitable for protecting against intentional alteration of data. Firstly, as there is no authentication, the data can be altered and CRC is computed without the substitution being detected [13]. Secondly, unsuspicious data can be hidden in the CRC of other data [16]. Due to this vulnerability, this method could not be used for proposed system.

4) Integrity through Hashing Techniques

In Hashing is a cryptographic function that takes a set of data as input and produces a thumbprint of this data, which is called a message digest. It maps large data sets of variable length, called keys, to smaller data sets of a fixed length. Most Popular Algorithms are: MD5 (16 Byte hash result), SHA1 (20 Byte hash result), SHA256 (32 Byte hash result) and SHA512 (64 Byte hash result).

In the proposed framework the two hash algorithm MD5 and SHA512 are used for experiment which verified both integrity texts as well as web images of web page convincingly. First MD5 hash algorithm [17] is implemented to check data integrity then SHA512 hashing [19] is implemented. SHA512 hash algorithm became the main technique for the proposed framework as it gave highest accuracy among the devised techniques illustrated earlier in this paper.

E. Web Content Repository

In this module repository is made to store featured values as a database for checking text and image defacement. Repository contains hash values, CRC32 value of defaced text and images and also contains PSNR and SSIM values which are calculated with the help of original and defaced images. The sample of repository made at the back end of the system is shown by Table III.

V. EXPERIMENTAL SETUP

The proposed work is conducted in the web security laboratory of MANIT, Bhopal. The lab server is used for the experiments along with one client machine. Hardware and software configurations of this server are: Intel X core CPU E7420 @ 2.13 GHz; 7 no. of CPU; 4GB RAM; CentOS Linux 5.9. This server is used to run sample websites and the client machine is used as a monitoring system.

To implement the methods CRC32 checksum, MD5, SHA512, PSNR and SSIM for the verification of integrity of text and web images, C#.Net language is used which runs in .NET framework environment. The software used for C#.Net is Microsoft Visual Studio 2010.

C# is used to building "real" and business applications. It has the ability to allow you to interoperate with any other language on the .NET platform.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

TABLE III. SSIM VALUES

UID	Page Name	Data Type	CRC32 Value	MD 5 Hash Value	SHA 512 Hash Value	PSNR Value	SSIM Value
G:\fd1	Webpage1 .htm	Image	5d59680d	fb7bfaac21ec f918fa56d82d	114b147a73572710981d4f35008 587b3f7487e916f8198022c303d 01a3e97d8415c0598bfd6399840	--	--
G:\fd2	Webpage2 .htm	Image	4d59680c	kc7bfaac21ec f918fa56d82d	114b147a73572710981d4f35008 587b3f7487e916f8198022c303d 01a3e97d8415c0598bfd6399840	18.09785	0.07645
G:\fd3	Webpage1 .htm	Image	785c4a29	bcf378c60f21 0149bea6ce4	becd520f2c2018d204381be5e23 c8729c321b825281d9f040b5344 1e000d9078c067180637da2893f	--	--
G:\fd4	Webpage2 .htm	Image	685c4a23	bcf378c60f21 0149bea6ce4	becd520f2c2018d204381be5e23 c8729c321b825281d9f040b5344 1e000d9078c067180637da2893f	∞	1
G:\fd5	Webpage1 .htm	file1.txt	785c4a29	67ba47f40cd 7d9cabddb64	597f5abd3ed253ee1204a80ab2e b5ab1dcebbcf946e702c3035e34 a1c1ca91a34440ddf74556e8d68	--	--
G:\fd6	Webpage2 .htm	file1.txt	785c4a29	a443ec6eab1a de2f0c2ab517	edc6c005c85abf45665967d216b e1535e3002bfed9f2fabc3286e31 9a1f7376b55f6834b9e316d983d	Not Defined	Not Defined

The sample websites containing more than 100 web pages are investigated. The used web pages contain the html tag and images such as JPEG, GIF and PNG.

As shown in Table IV, in training dataset out of 150 images, 120 images are tampered with 12 type's alteration (i.e. 10 images are chosen for each alteration) and out of 100 web pages 20 web pages are taken to alter text.

VI. EVALUATION

To evaluate the effectiveness of integrity methods, accuracy is calculated by measuring true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn). The terms *positive* and *negative* refer to the expected results and the terms *true* and *false* refer to observed results. The definitions of tp, tn, fp and fn for this proposed system are as follows:

- tp (correctly identified): It means image or text is *defaced* and also detected *defaced*.
- fp (incorrectly identified): It means image or text is *not defaced* but detected *defaced*.
- tn (correctly rejected): It means image or text is *defaced* but detected *not defaced*.
- fn (incorrectly rejected): It means image or text is *not defaced* and also detected *not defaced*.

Equation (7) is the formula to calculate accuracy which is given by:

$$Accuracy = \frac{tp + tn}{(tp + tn + fp + fn)} \quad (7)$$

Training data is taken to evaluate number of tp, tn, fp and fn (i.e. expected and observed results) during the experiment for CRC32, MD5, PRNR, SSIM and SHA512 techniques. Evaluation of accuracy for each method is demonstrated as follows: firstly, the methods to verify text and image integrity

TABLE IV. TRAINING DATASET FOR WEB IMAGES

Original Image		Tampered Image		Type of Alterations
Name	Size (in pixel)	Name	Size (in pixel)	
car.png	400 × 300	car.png	40 × 30	10x smaller
logo.png	20 × 15	logo.png	400 × 300	10x larger
mnit.png	400 × 300	mnit.png	200× 100	crop
abat.png	400 × 300	abat.jpg	400 × 300	Change type
wall.png	350 × 200	---	---	Text is inserted on image
sss.png	400 × 300	koala.jpg	400 × 300	Two images are merged
xdm.png	300 × 200	xdm.png	300 × 200	Blur
apj.png	500 × 200	apj.png	500 × 200	Rotate
f842.jpg	40 × 30	F842.jpg	40 × 30	Black & white
abc.png	80 × 60	abc.png	80 × 60	Same MSE value
dee.png	200 × 300	dee.png	200 × 300	Same image

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

are evaluated named as CRC32, MD5 and SHA 512. Then PSNR and SSIM methods are evaluated for image integrity verification. But PSNR, SSIM, CRC32 and MD5 failed to achieve required accuracy as compared to SHA 512.

In the proposed framework, SHA 512 technique is employed to reveal the website defacement because it gave maximum accuracy 96.80% among other methods named as CRC32, MD5, PSNR and SSIM. The accuracy is calculated by using the formula given in (7) and with the help of Table V.

TABLE V. EVALUATION FOR IMPLEMENTED INTEGRITY METHODS

	SSIM	PSNR	CRC 32	MD 5	SHA 512
Total Tested Sample	150	150	250	250	250
tp	29	30	35	37	22
fp	05	05	00	00	00
tn	98	100	190	190	220
fn	18	15	25	13	08
Accuracy (In %)	**84.66**	**86.66**	**90.00**	**94.80**	**96.80**

Experimental evaluation shows that SSIM has lowest accuracy. Different reasons behind lesser accuracy of CRC32, MD5, PSNR and SSIM are as follows: in CRC32 data can be hidden of other CRC32 code and it is the least secure algorithm whereas this data hiding is not possible with SHA512. MD5 has a collision problem [18] but SHA 512 does not have collision problem because this produces a 512-bit message digest which is roughly four times larger than that of MD5. PSNR and SSIM do not work for text integrity verification and SHA512 technique works for both integrity verification text as well as images.

On the basis of evaluated results from implemented integrity techniques a comparative graph is plotted shown in (5). In this figure integrity methods are put across x-axis of graph and values of test outcomes i.e. tp, tn, fn and fp are expressed on y-axis.

It is clear by the graph that amongst five techniques used; SHA 512 integrity technique shows highest accuracy.

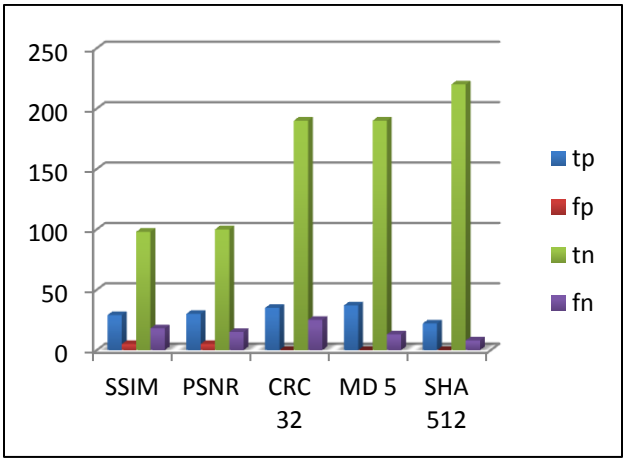


Figure 5. Comparisons of Implemented Integrity Techniques

VII. CONCLUSION

To accurate scrutinizing and checking the integrity of web content is still the challenge for law enforcement agencies. To handle the web defacement cases an effective prototype system has been developed which successfully point out the suspicious activity. The integrity of web contents is checked through CRC32, MD5, SHA512, PSNR and SSIM techniques. The best result is obtained by enforcing SHA512 technique. The work presented in this paper is focused on effective extraction and matching of web content. In future, further enhancements may be carried out, by devising an effective continuous polling mechanism in monitoring system.

REFERENCES

[1] Yu, W.D.; Nargundkar, S. and Tiruthani, N.; "A phishing vulnerability analysis of web based systems," in Computers and Communication, 2008. ISCC 2008. IEEE Symposium., Marrakech, 2008, pp. 326 - 331

[2] Shar, L.K. and Hee Beng Kuan Tan, "Defeating SQL Injection," in IEEE Computer, Singapore, 2013, Vol. 46, Issue: 3, pp. 68-77.

[3] Yusra A. Y. Al-Najjar and Der Chen Soong, "Comparison of Image Quality Assessment: PSNR, HVS, SSIM, UIQI," in International Journal of Scientific & Engineering Research, Vol. 3, Issue 8, August 2012.

[4] Alain Horé and Djemel Ziou, "Image quality metrics: PSNR vs. SSIM," in Pattern Recognition (ICPR), 2010 20th International Conference, 2010, pp. 2366 - 2369

[5] "Chinese websites 'defaced in Anonymous attack' ", [Online], Available: http://www.bbc.co.uk/news/ technology-17623939, April 5, 2012.

[6] Rick Burgess, "Hackers continue Israel attack, deface website, more credit cards leaked", Available: http://www.techspot.com/news/47046-hackers-continue-israel-attack-deface-website-more-credit-cards-leaked.html, January 13, 2012.

[7] "DOTC site defaced by 'Turkish' hackers", Available: http://www.gmanetwork.com/news/story/277286/scitech/technology/dotc-site-defaced-by-turkish-hackers, October, 2012.

[8] Bartoli, A.; Davanzo, G., and Medvet, E.; "A Framework for Large-Scale Detection of Web Site Defacements", ACM Transaction on Internet Technology (TOIT), Vol. 10, No. 3, Oct. 2010, Article 10.

[9] Kanti, T.; Richariya, V. and Richariya, V.; "Implementing a Web Browser with Web Defacement Detection Techniques", World of Computer Science and Information Technology Journal (WCSIT), Vol. 1, No. 7, 2011, pp. 307-310.

[10] Pan Shi; Xu, H. and Zhang, X. (Luke); "Informing Security Indicator Design in Web Browsers," in proceeding iConference '11 Proceedings of the 2011 iConference, Feb 11, 2011, pp. 569-575.

[11] Staikos, G. 2005. Web Browser Developers Work Together on Security. http://dot.kde.org/1132619164/, Nov. 2005.

[12] Moore, T. and Clayton, R.; "Examining the impact of website take-down on phishing," in Proceedings of the anti-phishing working groups' 2nd annual eCrime researchers summit, Pittsburgh, Pennsylvania, 2007, pp. 1-13.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[13] Peterson, W. W. and Brown, D. T. (January 1961). "Cyclic Codes for Error Detection," in Proceedings of the IRE 49 (1), 1961, pp. 228–235.

[14] Teng Lv and Ping Yan, "A Web Security Solution based on XML Technology," in Communication Technology, 2006. ICCT '06. Int. Conf., 2006, pp. 1 – 4.

[15] Suresh, R.M. and Padmajavalli, R., "An Overview of Data Preprocessing in Data and Web Usage Mining," in Digital Information Management, 2006 1st International Conference, Bangalore, India, 2006, pp. 193 – 198.

[16] Stigge, M. et al., "Reversing CRC – Theory and Practice" Berlin: Humboldt University Berlin Public Rep., May 2006.

[17] Rivest, R.; "The MD5 Message-Digest Algorithm," RFC 1321, Available: http://tools.ietf.org/html/rfc1321.

[18] Stevens, M.M.J.; "On Collisions for MD5", Master's thesis, Mathematics and Computing Science Department, Eindhoven University of Technology, Eindhoven, June 2007.

[19] U.S. National Security Agency, "Secure Hash Standard," in Federal Information Processing Standards Publication 180-2, Aug 2002.

[20] Ndajah, P. et al, "SSIM Image Quality Metric for Denoised Images," in Proc. 3rd WSEAS International Conference on Visualization, imaging and simulation, Japan, September 3, 2010, pp. 53-57.

[21] Wang, Z.; Simoncelli, E.P. and Bovik, A. C.; "Multi-Scalestructural Similarity For Image Quality Assessment," in Proc. 37th IEEE Asilomar Conference on Signals, System and Computation, Pacific Grove , CA , Nov 2009-10.

[22] Huynh-Thu, Q.; Ghanbari, M.; "Scope of validity of PSNR in image/video quality assessment," in Electronics Letters (Volume: 44, Issue: 13), June 19 2008, pp. 800 – 801.

[23] MIT.edu, Available: http://people.xiph.org/~xiphmont/ demo/theora/demo7.html.

[24] Oriani, Emanuele. "qpsnr: A quick PSNR/SSIM analyzer for Linux", April 2011, Available: http:// qpsnr. youlink.org/.

[25] O"Opinion Poll", Available: http://en.wikipedia.org/wiki/ Opinion_poll, April 17, 2013.

Java Card For PayTV Application

Pallab Dutta
Team Leader, Centre For Development Of Telematics
Electronic City, Bangalore, India.

***Abstract*— Smart cards are widely used along with PayTV receivers to store secret user keys and to perform security functions to prevent any unauthorized viewing of PayTV channels. Java Card technology enables programs written in the Java programming language to run on smart cards. Smart cards represent one of the smallest computing platforms in use today. The memory configuration of a smart card are of the order of 4K of RAM, 72K of EEPROM, and 24K of ROM. Using Java card provides advantages to the industry in terms of ease of coding, faster time to market and faster upgrades as compared to plain smart cards . Also different applications like payTV, e-commerce, health-card can easily be implemented in a single java card as multiple applets corresponding to each application can coexists in a single java card. But there are security concerns in java cards and also the performance issues. In this paper, we analyse the suitability of using Java card for PayTV applications as part of conditional access system in place of plain smart cards.**

Keywords- Smart Card, Java Card, PayTV, Conditional Access System (CAS), Cryptography

I. INTRODUCTION

Satellite PayTV network consists of a head-end, uplinking transmission system, satellite and reception system along with subscriber Management system (SMS) and subscriber Authorization System (SAS). At the head-end (transmit side), the digital content (including video, audio and data), which the operator wishes to restrict access, is scrambled (DVB-CSA) by the control word (CW) derived from a constantly changing pseudo-random binary sequence generator. The control word also needs to be protected: the control word is encrypted with a service key (also called authorization key) (SK / AK) . The encrypted control word is then packaged into entitlement control message (ECM). Then, the service key is encrypted with the individual key (IK) supplied by the subscriber management system (SMS) and is then packaged with entitlement data into entitlement management message (EMM). Finally, the scrambled content, entitlement control message, and entitlement management message are together broadcasted in the same channel. This is shown as in Figure 1.

At the reception-end, the set-top box (STB) and smart card authenticate each other. STB filters entitlement management message and entitlement control message according to the parameters provided by the smart card (SC) and then forwards these messages to smart card. Smart card decrypts entitlement management message using individual key stored in smart card to get service key/authorization key and the entitlement data. After having passed the verification of the access entitlement, smart card uses the service key to decrypt the encrypted control word and returns the control word towards set-top box so that set-top box will be allowed to descramble the scrambled content. This is shown in the Figure 2. The smart card and part of STB where entitlement messages are filtered and smart card handshake takes place is called conditional access sub system.

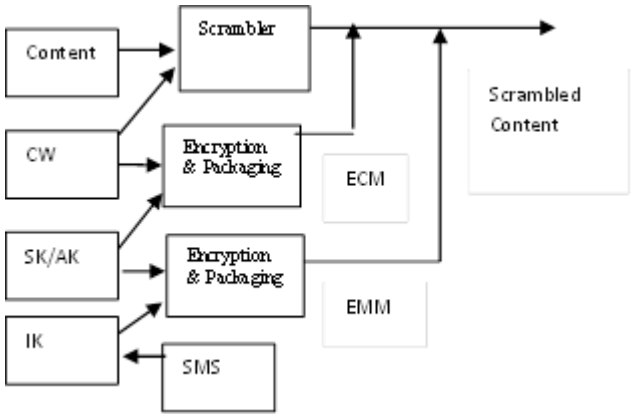


Figure 1. PayTV Headend

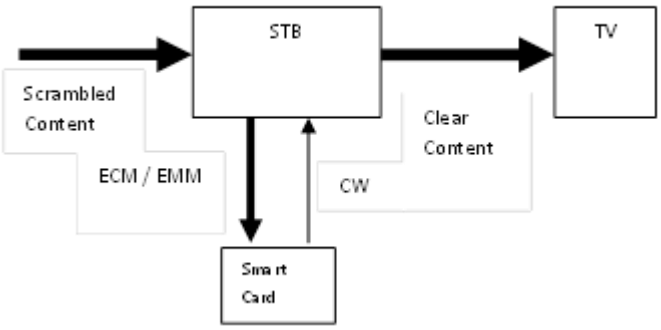


Figure 2. PayTV Receiver

Java Card technology enables programs written in the Java programming language to run on smart cards. A JavaCard is a typical smart card: it conforms to all smart card standards and thus requires no change to existing smart card-aware applications. However, JavaCard has a specific important feature that makes it unique: a Java Virtual Machine is

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

implemented in its read-only memory (ROM) mask. The JVM controls the access to all smart card resources, such as memory and I/O, and thus essentially serves as the smart card's operating system. The JVM executes a Java bytecode subset on the smart card, ultimately providing the functions accessible from outside, such as signature, log-in, and loyalty applications. Java Card offers definite advantages over plain Smart cards in terms of ease of programming, faster time to market, coexistence of multiple applications, interoperability etc. In this paper we analyse the suitability of using java card for PayTV applications as part of conditional access system in place of plain smart cards. In Section II, we list the advantages and disadvantages of Java Card over conventional Smart card. Section III analyses java card technology in detail from architecture and security point of view. In Section IV, we give the requirements to be met by java card for using in PayTV application. In Section V, we have given the experimental setup and performance details of java card. Section VI, gives the conclusion.

II. ADVANTAGES AND DISADVANTAGES OF JAVA CARD

A. Advantages of Java Card

There are several advantages of java card over plain smart cards. The following are the main advantages of Java Cards:

Interoperable: Applets developed with Java Card technology runs on any Java Card technology-based smart card, independently of the card vendor and underlying hardware.

Secure: Java Card technology relies on the inherent security of the Java programming language to provide a secure execution environment. It was designed through an open process, and the platform's proven industry deployments and security evaluations ensure that card issuers benefit from the most capable and secure state of the art technology.

Multi-Application Capable**:** Java Card technology enables multiple applications to co-exist securely on a single smart card. Hence same card can be used for several end applications.

Dynamic: New applications can be installed securely after a card has been issued, providing card issuers with the ability to dynamically respond to their customer's changing needs.

Open**:** Java Card application developers benefit from object-oriented programming and design, and have access to off-the-shelf Java development tools.

Faster Development: Because of Java based development environment, this gives the advantage of faster time to market and also debugging easier and more error free code can be developed faster compared to a native application development.

Compatible with Existing Standards: The Java Card API is compatible with international standards for smart cards such as ISO7816, or EMV. It is referenced by major industry-specific standards such as Global Platform and ETSI.

B. Disadvantages of Java Card

Although java card offers lot of advantages, it also has a few disadvantages; following are the main disadvantages of Java card:

Less Memory Space: Java Card Run time Environment occupies extra memory space in the java card, hence the memory space available for the user application reduces.

Slower: Java code execution is normally slower than the native code written in middle level or low level programming languages.

Security: Because java card is an open system, security through obscurity does not workout.

III. JAVA CARD TECHNOLOGY AND SECURITY

A. Java Card Technology

Before going into security issues let us first discuss about Java Card Architecture.

As shown in the figure-3, the low level operations like physical data transfer, file management, communication and instruction execution are handled by microcontroller program and operating system.

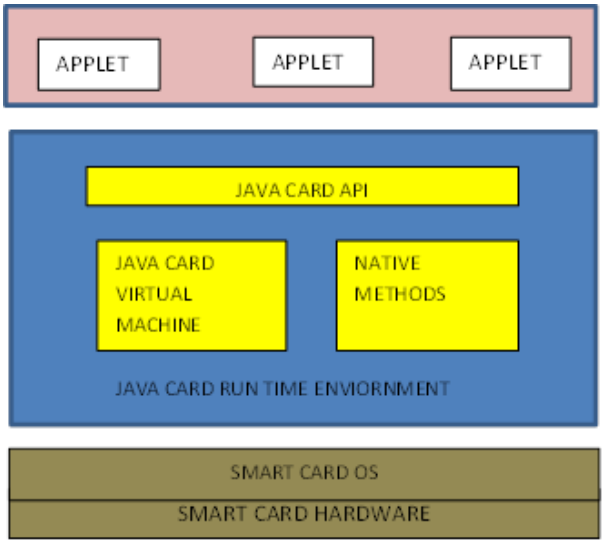


Figure 3. JAVA Card Architecture

JCRE (Java Card Runtime Environment) sits over this layer and it consists of Java Card API, JCVM (Java Card Virtual Machine) and native methods. The native methods are platform dependent and they directly interact with underlying operating system, it provides services like cryptography,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

memory allocation and I/O operations. Java Card API also called as framework consists of all Java packages required for various Java Card functionalities. JCVM executes the Java Byte Code just like JVM but the difference is that JCVM is split into two parts. One is Java Card Converter which is off-card, converts *.class* files into card compatible format (*.cap* files) and the other is on-card part which interprets the byte code. JCRE and operating system program are stored into ROM at the time of manufacture of chip while card applications are loaded into EEPROM.

Java Card Language Subset: Because of its small memory footprint, the Java Card platform supports only a carefully chosen, customized subset of the features of the Java language. This subset includes features that are well suited for writing programs for smart cards and other small devices while preserving the object-oriented capabilities of the Java programming language. Table 1 highlights some notable supported and unsupported Java language features. Keywords of the unsupported features are also omitted from the language. Many advanced Java smart cards provide a garbage collection mechanism to enable object deletion.

TABLE 1: Supported and Unsupported Java Card features

Supported Java feature	Unsupported Java Feature
• Small primitive data types: boolean, byte, short • One-dimensional arrays • Java packages, classes, interfaces, and exceptions • Java object-oriented features: inheritance,virtual methods, overloading and dynamic object creation, access scope, and binding rules • The int keyword and 32-bit integer data type support are optional.	• Large primitive data types: long, double, float • Characters and strings • Multidimensional arrays • Dynamic class loading • Security manager • Garbage collection and finalization • Threads • Object serialization • Object cloning

Java Card Application Deployment: Java card application is first compiled using Java compiler and class files are obtained. These class files are converted into cap files (Converted Applet files) and then installed on to the card as shown in the figure 4. Now applet must be installed and registered in the card. This is done either by the installation program sitting on JCRE or by the *install* method of abstract applet class. When the installation and registration is successful, an applet is instanced and all objects related to it are created.

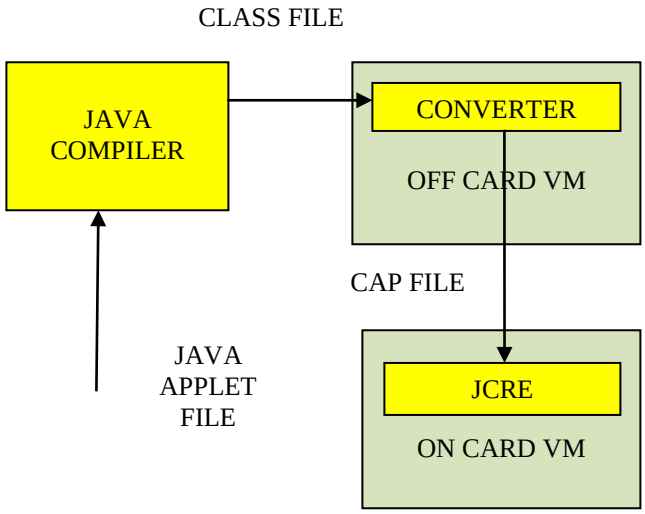


Figure 4. Java Card Application Deployment

B. *Java Card Security*

1. *Java card security features:*

Java Card provides lot of security features like applet firewalls, Java type safe, object sharing etc. As smart card uses cryptography technique for data encryption, Java Card provides Crypto API. We discuss about the features in detail below:

Prohibition of Dynamic Class Loading: Java Card does not allow dynamic class loading. This means that Java classes are not be downloaded on-the-fly during execution; a security risk assumed by standard Java. Instead, Java applets are usually burned in during masking when the card is initialized and personalized prior to being fielded.

Applet Firewall: Java Card can handle multiple applets and these applets can be from different venders so it is not secured if objects of one applet are accessible to another applet. Java Card uses applet firewall and type safe feature of Java to restrict unauthorized access of data. That is Java Card defines context for each applet which represents all the objects accessible to a particular applet.

Object Sharing: Java Card also provides a facility to share objects by implementing *Sharable* interface. This allows the applets to share objects and at the same time it restricts them from accessing methods which are accessible only to the owner of the object that is remote method invocation is not allowed.

Java Card API: The Java Card APIs consist of a set of customized classes for programming smart card applications according to the ISO 7816 model. The APIs contain three core packages and one extension package. The three core packages are java.lang, javacard.framework, and javacard.security. The

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

extension package is javacardx.crypto. Many Java platform classes are not supported in the Java Card APIs. For example, the Java platform classes for GUI interfaces, network I/O, and desktop file system I/O are not supported. The reason is that smart cards do not have a display, and they use a different network protocol and file system structure. Also, many Java platform utility classes are not supported, to meet the strict memory requirements. The classes in the Java Card APIs are compact. They include classes adapted from the Java platform for providing Java language support and cryptographic services. They also contain classes created especially for supporting the smart card ISO 7816 standard.

2. *Security Risks in Java Card*

Although java card offers lot of security features, it also has few security risks. The following are the main risks associated with java cards.

Absence of Garbage Collection: In Java, Garbage Collection feature controls the freed and unused objects of the applet but this feature is absent in Java Card (because of its memory and processor resource limitations) which may cause serious threat like handle of a free object is allocated to a new object. If any error occurs in the program then the memory allocated to the object is never freed which lead to memory leakage. This is a serious problem especially on smart card with such limited memory resources and application may get hanged up in the middle of transaction and suffer from denial-of-service attack**.**

Native methods: JCRE interacts with card resources using native methods which are mostly written in C or C++, card vender can upgrade card by adding new native methods and applets can use these methods. Java card firewalls have no control over these native methods so all the firewall security fails when it comes to native methods.

Exception Handling: Another security threat is with exception handling, it can be controlled if applets are intensively tested using testing tools before loading on card. If an exception is not handled in any of the called methods it becomes unpredictable and may hang the applet and corrupt the card.

Protocol Interactions: In today's smart card market, multi-application cards are becoming very common and also convenient. That is, rather then carrying several different cards for different purposes, the applications are integrated on a single card. For example, a smart card may provide driver's license or social security identification, a credit line, a debit account, e-cash, health insurance information, calling card applications, digital certificates for authentication, and different loyalty programs. Insofar as multiple applications implement different protocols (many of which may require authentication), unintended protocol interactions may compromise the security of the card and its multiple applications. Because of small memory constraints in smart cards, it is currently impractical to store unique cryptographic public/private key pairs for each different application requiring cryptographic functions. Because of this, multiple applications often share the same key material. Different protocols that share the same key material can introduce potential security problems that would not otherwise exist with each protocol considered in isolation. That is, the security of two different protocols do not necessarily compose, especially given protocol interactions.

Physical Security: The physical security risks of Java cards are same as those of plain smart cards. Various physical attacks like, DPA (differential Power Analysis) need to be countered measured in Java Cards in a similar way that of plain smart cards.

IV. REQUIREMENTS TO BE MET BY JAVA CARDS FOR USING IN PAYTV APPLICATION

A. *Security of Cryptographic Algorithms*

In PayTV applications, both symmetric and asymmetric algorithms are used. Although, PayTV standards does not mandate usage of any specific cryptographic algorithm for key distribution and scrambling, but generally two layer security scheme is employed as mentioned in the section I of this paper. Symmetric cryptographic algorithm like AES and Triple DES are used. An asymmetric algorithm like RSA is used. Pseudo Random numbers are required to be generated as part of authentication protocol used as session keys between smart card and STB. The implementation of complex crypto algorithm shall be easier and faster through standard APIs.

B. *Confirmation to ISO7816 standard*

The Smart Card – PayTV receiver interface is ISO 7816, hence the Java card shall also confirm to this specification. The Java Card APIs consist of a set of customized classes for programming smart card applications according to the ISO 7816 model. The classes in the Java Card APIs are compact. They include classes adapted from the Java platform for providing Java language support and cryptographic services. They also contain classes created especially for supporting the smart card ISO 7816 standard.

C. *High Security:*

In the payTV scenario, if the control word (CW) is cracked, then the TV channels can be decoded unauthorised. This will lead to huge revenue loss to the operator. The Smart Card and the receiver – any one of these two may be malicious; hence requires bi-directional authentication so that the receiver is able to verify the identity of the Smart Card and also Smart Card is able to verify the identity of the receiver. Otherwise, a cloned smart card will be able to decode the CW or a fake receiver will be able to store secret information from a genuine Smart Card and create large number of fake Smart Cards. It

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

shall be possible to implement bi-directional authentication in Java Card in an efficient way both in terms of memory usage and time required for executing complex crypto algorithm.

D. Performance

The Java card to be used in PayTV receiver need to satisfy stringent performance requirements. In payTV system, Control word usually changes in every 10 seconds, also control word is per channel. For better user experience, ECM is transmitted every 100ms, so Smart card need to decode the ECM message and extract Control Word and send it to the receiver via ISO 7816 interface, once in every 10s minimum. Generally Control Word extraction and transmitting it to the receiver involves triple DES or AES encryption / decryption. When the receiver is switched on or Smart Card is inserted in the receiver, bi-directional authentication takes place between the smart card and the receiver. This involves RSA decryption, Random Number Generation and AES decryption. This phase has to be completed within 2-3 seconds for smooth user experience.

V. EXPERIMENTAL SETUP AND PERFORMANCE

The experimental setup consists of a core2 duo PC acting as PayTV receiver, a smart card reader is connected to the USB port of the PC. JCOP 21 Java card is used to perform the cryptographic functions triple DES, AES, RSA. Eclipse IDE is used along with JCOP plugins for the Java Card Applet Development. The Java card used here has triple DES, AES and RSA coprocessors. The EEPROM size is 72K and RAM size is 4096 B. Applets are loaded in EEPROM. 128 bit key length for AES and 1024 bit key of RSA is used.

It is observed that Triple DES encryption is taking around 5 ms, AES is taking around around 10 ms. RSA takes around 100 ms. As we have seen in the performance requirements, the CW changes in every 10 second and ECM is repeated per 100ms, hence the Triple DES and AES performances are acceptable for PayTV application. Also RSA is not used runtime for CW extraction and RSA is used only when smart card is inserted in the receiver or when the receiver is turned on with the smart card, hence the RSA performance is also acceptable for the given application of payTV conditional access system.

VI. CONCLUSIONS

We have discussed the payTV conditional access system requirements and analysed the advantages and disadvantages of Javacard in general and also in the specific context of payTV receiver. We have seen that although there are some disadvantages of Java Cards but those do not create any hindrance for the Java Cards to be used in the payTV receivers as part of conditional access system. As per Kerckhoff's principle, an well-known principle in cryptography, security though obscurity does not give overall system security in long run, hence an open smart card platform like Java Card platform, does not compromise the security of a proprietary smart card system rather it may enhance the security. Also as we have seen from the basic performance data, the triple DES, AES and RSA gives the acceptable performance for payTV applications. However in some practical payTV systems, there will be additional computational load wrt certificate verification for mutual authentication, dynamic session key generation etc., and in cases, it is required to implement Schnorr Algorithm. This will require modulo multiplication operation and not be possible to use the RSA co-processor directly. In such cases a proper mechanism to use the individual building blocks of the co-processor need to be designed and made known to the Java Card user to use the computational power of the co-processor to achieve acceptable performances in such cases also.

REFERENCES

[1] Wolfgang Rankl, Smart Card Applications, Wiley, 2007.

[2] Drioscoll Gerardo, Essential guide to Digital Set top boxes and interactive TV, Prentice Hall, 2000.

[3] Zhiqun Chen, Java Card Technology for Smart Cards: Architecture and Programmer's Guide published by Addison Wesley, June 2000.

[4] Surender Reddy Adavalli, "Smart Card Solution: Highly secured Java Card Technology", www.cs.auckland.ac.nz /courses/compsci725s2c/.../725adavalli.pdf.

[5] Anup K. Ghosh: Security Risks of Java Cards, In Proceedings of the Twelfth IFIP WG 11.3 Working Conference on Database Security, Greece, 1999.

[6] Liu Yongliang, Xiaolin Yang, Hongxun Yao, and Wen Gao, "Novel Secure Communication Protocol for Conditional Access System", International Journal of Network Security, Vol.5, No.2, Sept 2007, PP.121-127

[7] Tianpu Jiang, Yongmin Hou and Shibao Zheng, "Secure Communication between Set-top Box and Smart Card in DTV Broadcasting", IEEE transaction on Consumer Electronics, Vol50, No. 3, AUGUST 2004, PP. 882-886.

[8] Eun-Jun YOON and Kee-Young YOO, "Robust Key Exchange Protocol between Set-Top Box and Smart Card in DTV Broadcasting", INFORMATICA, 2009, Vol. 20, No. 1, PP 139-150.

[9] T. Jiang et al, "Key distribution based on hierarchical access control for Conditional Access System in DTV broadcast," IEEE Trans. on Consumer Electronics, Vol. 50. Feb. 2004, pp.225-230.

[10] C. P. Schnorr, "Efficient identification and signatures for smart cards," LNCS, Vol 435/1990, PP. 239-252.

[11] Hyo Kim, "Secure Communication in Digital TV Braodcasting", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.9, September 2008.

[12] ISO/IEC 13818-1: Information technology –Generic coding of moving pictures and associated audio information, International Organization for Standardization, 2000.

[13] ISO/IEC 7816-3: 1997, Information technology – Identification cards – Integrated circuit(s) cards with contacts – Part 3: electronic signals and transmission protocols, International Organization for Standardization, 1997.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[14] ISO/IEC 7816-4: Information technology – Identification cards – Integrated circuit(s) cards with contacts – Part 4: Inter industry commands for interchange, International Organization for Standardization, 1997.

AUTHORS PROFILE

Mr.Pallab Dutta is working in C-DOT (Centre For Development Of Telematics) for last 16 years. He has worked in various fields/projects of Telecommunication during this tenure. His area of interests are Embedded system design, High Performance Computing, Cryptography, Multi-Agent system etc. He Did BTech from REC (presently known as NIT) Calicut in ECE in 1997. He completed his ME (CSE) in 2007 and presently pursuing his PhD in CSE.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IMPLEMENTATION OF OMNI DIRECTIONAL COGNITIVE MEMORY IN OFFICE MOBILE ROBOT

S.Sivagnana sundari,
Research Scholar, Department of Statistics,
Manonmaniam Sundaranar University,
Tirunelveli, INDIA,

Dr.C.Vijayalakshmi,
Professor, School of Advance Sciences
Department of Mathematics Division,,
VIT University, Chennai, INDIA,

***Abstract*-This paper presents a method for building omni directional memory from two dimensional memory storage. The omni directional memory is implemented in the office mobile robot. Time series method is used to estimate the next position of the robot based on the stored memory. Images and sounds are collected in the office environment to store expert database in the memory of the robot. A section of the image frames taken in the corridor and how the image is associated in the omni directional memory is shown. Based on the information in the memory and the speed with which the robot is moving, the method of predicting the next position by time series method is discussed.**

Keywords: Omni directional, Mobile robot, Two dimensional memory, Time series method

I INTRODUCTION

The concept of omni directional memory plays an important role in implementing a full fledged self navigating autonomous office mobile robot. Conventional robots are mostly fixed to a point or they are guided with remote control or through wired path. Heavy operations require robot to be mostly fixed to a place. A preprogrammed movements will be stored using teach pendent procedures. However, robots that are on the mobility should be able to navigate. The robot must be able to retrace and come back to the starting point without any guidance of floor pasted line sensors. For the robot to move from starting point to the required destination and back to starting, information like images and sounds should be preprogrammed in one form or the other. The information for preprogramming can be any one or combinations of the following:

i) By knowing the distance of important landmarks from the starting point to reach a given destination.

ii) By knowing the distance between adjacent landmarks.

iii) In case any one of the landmarks has been displaced, then distance between present adjacent landmarks should be known.

iv) By using the images acquired at different positions on route from starting point to destination and back to starting point.

v) If the robot is displaced on its route, but still it has to find way to reach next landmark.

vi) By using the sounds that occur regularly to confirm the path.

II .LITERATURE REVIEW

Cao zuo et al, 1986, described algorithm for omnidirectional vision navigation. A prototype omnidirectional vision system and the implementation of the navigation techniques using this modern sensor and an advanced automatic image processor is described.

Krotkov, 1989, assumed a mobile robot is equipped with a single camera and a map marking the positions in its environment of landmarks. The robot moves on a flat surface, acquires one image, extracts vertical edges from it, and computes the directions to visible landmarks. The problem is to determine the robot's position and orientation (pose) by establishing the correspondence between landmark directions and points in the map. This approach capitalizes on the excellent angular resolution of standard CCD cameras, while avoiding the feature-correspondence and 3D reconstruction problems. The problem is formulated as a search in a tree of interpretations (pairings of landmark directions and landmark points), and an algorithm to search the tree efficiently to determine the solution poses(s) is developed, taking into account errors in the landmark directions extracted by image processing. Quantitative results from simulations and experiments with real imagery are presented.

Leonard et al, 1992 presented an algorithm for autonomous map building and maintenance for a mobile

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

robot. They represent each feature in the map by a location estimate (the feature state vector) and two distinct measures of uncertainty: a covariance matrix to represent uncertainty in feature location, and a credibility measure to represent in the validity of the feature. During each position update cycle, predicted measurements are generated for each geometric feature in the map and compared with actual sensor observations. Successful matches cause a feature's credibility to be increased. Unpredicted observations are used to initialize new geometric features, while unobserved predictions result in a geometric feature's credibility being decreased. They describe experimental results obtained with the algorithm that demonstrate successful map building using real sonar data. Mobile robotic devices hold great promise for a variety of applications in industry. A key step in the design of a mobile robot is to determine the navigation method for mobility control.

Ativa and Hager, 1993, described an algorithm for determining robot location from visual landmarks. This algorithm determines both the correspondence between observed landmarks (in this case vertical edges in the environment) and a stored map, and computes the location of the robot using those correspondences. The primary advantages of this algorithm are its use of a single geometric tolerance to describe observation error, its ability to recognize ambiguous sets of correspondences, its ability to compute bounds on the error in localization, and fast execution. The algorithm has been implemented and tested on a mobile robot system. In several hundred trials it has never failed, and computes location accurate to within a centimeter in less than 0.5s

Yagi et al, 1994, described a conic projection image sensor (COPIS) and its application: navigating a mobile robot in a manner that avoids collisions with objects approaching from any direction. The COPIS system acquires an omnidirectional view around the robot, in real-time, with use of a conic mirror. Based on the assumption of constant linear motion of the robot and objects, the objects moving along collision paths are detected by monitoring azimuth changes. Confronted with such objects, the robot changes velocity to avoid collision and determines locations and velocities.

Yagi et al, 1995, designed an omnidirectional image sensor COPIS (Conic Projection Image Sensor) to guide the navigation of a mobile robot. The feature of COPIS is passive sensing of the omnidirectional image of the environment, in real-time (at the frame rate of a TV camera), using a conic mirror. COPIS is a suitable sensor for visual navigation in a real world environment. They report here a method for navigating a robot by detecting the azimuth of each object in the omnidirectional image. The azimuth is matched with the given environmental map. The robot can precisely estimate its own location and motion (the velocity of the robot) because COPIS observes a 360° view around the robot, even when all edges are not extracted correctly from the omnidirectional image. The robot can avoid colliding against unknown obstacles and estimate locations by detecting azimuth changes, while moving about in the environment. Under the assumption of the known motion of the robot, an environmental map of an indoor scene is generated by monitoring azimuth change in the image.

Branca et al, 1997, estimate the egomotion from the motion field to determine the robot's direction and the time to collide (TTC) with the environment. The mobile robot is restricted to travel on a flat surface in a stationary environment.

Development of sensors for small size indoor mobile robots have been proposed with sonar-ring which can measure accurate direction and distance to reflecting points rapidly for this purpose. This sonar ring needs to receive echo signal by multiple receivers and to detect multiple Time-of-Flights (TOFs) in received signal at each receiver simultaneously for achieving fast and accurate measurement, Teruko et al, 2000.

Stoffler and Schnepf,1998; Stoffler and Schnepf 2000, created an application for a mobile robot that is capable of detecting potential obstacles, roughly estimating their positions, and planning an avoidance task.

Krishna and Kalra, 2001, dealt with the advantages of incorporating cognition and remembrance capabilities in a sensor-based real-time navigation algorithm. The specific features of the algorithm apart from real-time collision avoidance include spatial comprehension of the local scenario of the robot, remembrance and recollection of such comprehended scenarios and temporal correlation of similar scenarios witnessed during different instants of navigation. These features enhance the robot's performance by providing for a memory-based reasoning whereby the robot's forthcoming decisions are also affected by its previous experiences during the navigation apart from the current range inputs. The environment of the robot is modeled by classifying temporal sequences of spatial sensory patterns. A fuzzy classification scheme coupled to Kohonen's self-organizing map and fuzzy ART network determines this classification. A detailed comparison of the present method with other recent approaches in the specific case of local minimum detection and avoidance is also presented. As for escaping the local minimum barrier is concerned this paper divulges a new system of rules that lead to shorter paths than the other methods. The method has been tested in concave, maze-like, unstructured and altered environments and its efficacy established.

The supervisory control of mobile robots using a biologically inspired short-term memory structure called a sensory Egosphere has been described. The Egospehere is implemented as a virtual geodesic dome upon which sensory data from the surroundings of the robot are written. The Egosphere is managed by a distinct agent in a distributed agent-based robot control architecture called the Intelligent Machine Architecture. A human-robot interface and a test bed for evaluating the control system has been proposed, Kawamura et al, 2001.

Memory-based methods estimate self-location by directly comparing the input image with stored images in memory. This approach has received much attention,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

because the localization accuracy will increase by simply increasing the number of images to memorize. In this paper, we propose a novel self-localization method based on eigenspace analysis as one of memory-based methods. We employ an omni directional image sensor that can take an image of surrounding environment of the sensor at video-rate. The proposed method identifies the sensor's location by evaluating the similarity between autocorrelation images which are converted from omni directional images and are invariant against the rotation of the sensor. To reduce the computational cost and memory space, autocorrelation images are projected into eigenspaces. The similarity is evaluated in eigenspaces to reduce the computational cost. The experimental results show that the method is capable of estimating the location of sensor even with low dimensional images, Nobuhiro et al, 2003.

The omni directional vision sensor for the view-based navigation method has been proposed with an extended model of a route called the 'omni view sequence'. The author propose a map named the 'view-sequenced map' which represents an entire corridor environment in a building. A method for the automatic acquisition of a view-sequenced map based on the exploration in a corridor using both stereo and omni directional vision is also described. The experimental results of the autonomous navigation and the map acquisition are presented to show the feasibility of the proposed methods, Yoshio et al, 2003.

A technique for vision-based robot navigation, Emanuele et al, 2004, with basic framework is to localize the robot by comparing images taken at its current location with reference images stored in its memory. The only sensor mounted on the robot is an omni directional camera. The Fourier components of the omni directional image provide a signature for the views acquired by the robot and can be used to simplify the solution to the robot navigation problem. The proposed system can calculate the robot position with variable accuracy ("hierarchical localization") saving computational time when the robot does not need a precise localization (e.g. when it is traveling through a clear space). In addition, the system is able to self-organize its visual memory of the environment. The self-organization of visual memory is essential to realize a fully autonomous robot that is able to navigate in an unexplored environment. Experimental evidence of the robustness of this system is given in unmodified office environments.

Campbell et al. 2004,2005, proposed a new technique for estimating egomotion on which a video camera is mounted on a mobile robot so that the optical flow field is divided into two portions: the ground region and the sky region which are separated by a horizon line. The motion in the ground portion is used to compute the robot translation while the sky portion is used to compute the robot rotation. The application is created for the mobile robot's visual odometry.

A method has been developed for a cognitive robot behavior, Palis et al, 2005, control in which a small number of behaviors are loaded into a workspace, called working memory, where they are combined to generate actions during a task execution. The existing components can be used i cognitive robot architecture, such as the Long-Term Memory, the Short-Term Memory, with the addition of a working memory system and a control mechanism called the Central Executive Agent to create a modular control system. This control method is used to drive the behaviors of our humanoid robot ISAC.

2005 The design and creation of an episodic memory system for a cognitive robot is presented. This memory system is interfaced with a machine emotion system with the goal of producing intelligent behaviors in a cognitive humanoid robot. The design of the system is tested and analyzed through the explanation of a case in which emotion assists in the retrieval of an episode, Will Dodd et al, 2005.

Briggs et al, 2006, presented a method for navigation and localization of a mobile robot equipped with an omnidirectional camera. They represent the environment using a collection of one-dimensional panoramic images formed by averaging the center scanlines of a cylindrical view. Such 1D images can be stored and processed with few resources, allowing a fairly dense sampling of the environment. Image matching proceeds in real time using dynamic programming on scale-invariant features extracted from each circular view. By analyzing the shape of the matching curve, the relative orientation of pairs of views can be recovered and utilized for navigation. When navigating, the robot continually matches its current view against stored reference views taken from known locations, and determines its location and heading from the properties of the matching results. Experiments show that our method is robust to occlusion, repeating patterns, and lighting variations.

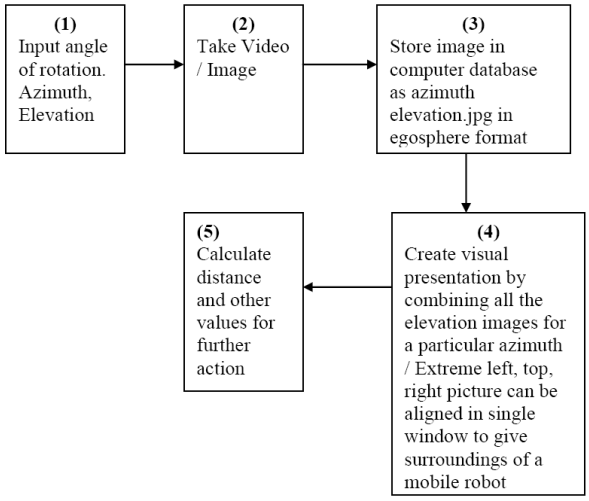
Ching-Chih et al, 2010, presented a behavior-based navigation method for a mobile service robot with a three-wheeled omnidirectional mobile platform. A heuristic fuzzy Kohonen clustering network (FKCN) is presented to obtain laser area weights and desired robot motion heading for all types of indoor environments. An efficient behavior-based navigation structure is then proposed to merge the heuristic FKCN with three fuzzy behaviors in order to navigate the robot without any collisions in crowded and cluttered indoor environments. Simulation and experimental results are conducted to verify the feasibility and effectiveness of the proposed method.

Jwu-Sheng Hu et al, 2011, proposed a framework that simultaneously localizes the mobile robot and multiple sound sources using a microphone array on the robot. First, an eigen structure-based generalized cross correlation method for estimating time delays between microphones under multi-source environments is described. A method to compute the far field source directions as well as the speed of sound using the estimated time delays is proposed. In addition, the correctness of the sound speed estimate is utilized to eliminate spurious sources, which greatly enhances the robustness of sound source detection. The arrival angles of the detected sound sources are used as observations in a bearings-only SLAM procedure. As the source signals are not persistent and there is no identification of the signal content, data association is

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

unknown which is solved using FastSLAM. The experimental results demonstrate the effectiveness of the proposed approaches.

III SCHEMATIC DIAGRAM



Give instruction of rotation through azimuth, elevation. Present position is known. If present position is not known for the machine vision system, then plan for home for the head to return to home position. The information is transmitted through wireless method. Information from the attached sensor is recorded, for example a proximity sensor will mention the distance between robot and the surroundings. Similarly an imaging sensor can describe about the lighting conditions. For each azimuth and elevation, store the information coming from each sensor with filenames as 'azimuth value' concatenated with elevation value concatenated with {image or distance or color}.

The image information can be taken at certain intervals. However, information from other sensors is to be analyzed and stored. Information received from each sensor at a particular distance and at a particular location is to be fused and presented. At certain points of time, it does not require receiving information from the camera; however by knowing the azimuth and elevation through the machine vision system along with the position of the robot, the relevant images can be retrieved based on the information stored. Based on the information received from the sensor, collaborative signal processing is done. Time series method is applied for next position of the Robot.

IV. TIME SERIES METHOD FOR ESTIMATING NEXT POSITION

A time series is a sequence of signals or observations $\mathbf{x}_t$, each one recorded at a specified time instant t. A discrete time series is one in which $\mathbf{x}_t$ are recorded at different time points. Without special notation, a discrete time series often refers to an observation sequence with fixed time interval between two neighboring observations. Our research handles only such discrete time series.

A time series is multivariate if there is more than one variable involved in each observation and if these variables are cross-related. The time series estimation method can be parametric and non-parametric. In parametric method, a model is assumed. However, there is no predefined model for non-parametric methods. Instead, features are extracted from the available time series data. The features can be mean values, variances, correlations and frequencies, and do the estimation job based on these features.

A parametric approach like AutoRegression Moving Average, ARMA (p, q) is proposed. ARMA (p, q) is used for modeling stationary and Gaussian-distributed time series.

ARMA (p, q) is an ideal model; for a certain time series sample, there is one and only one ARMA(p, q) model with specified coefficients corresponding to it. This property makes the estimation. Compared with non-parametric methods, our ARMA (p, q) model based estimation methods provide lag influence and interactions, which is not easily summarized by features.

Given a time series observation sequence, we can estimate the parameters of the ARMA model which generates this time series. Hence, representing a time series sequence by a limited dimensional vector consisting of the ARMA parameters is done.

ARMA (p,q) model

In time series, the current signal x_t influenced by the previous signals x_{t-1}, x_{t-2}, and so on. A general time series model may be formed as,

$$x_t = f(x_{t-1}, x_{t-2}, ..., u_{t-1}, u_{t-2}) + \xi_t$$

Where u_t is the control knob and ξ_t is the white noise at time t.

A linear model is tractable and is given by:

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + ... + \alpha_p x_{t-p} + \beta_1 \xi_{t-1} + \beta_2 \xi_{t-2} + ... + \beta_q \xi_{t-q} + \xi_t \quad (1)$$

.

Without the control term, the equation is called Auto Regression model, or AR (p), in which p is the window size.

The α s are the coefficients of the AutoRegression part, and β s are the coefficients of the Moving Average part.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Memory-based time series recognition

Memory-based time series estimation consists of following four phases.

Collection of time series samples

Many time series samples are collected (images and sounds from important locations of the path of the robot). The parameters $\hat{\alpha}$ and $\hat{\beta}$ are involved in the ARMA (p, q) model.

If several time series samples shares the same $\hat{\alpha}$ and $\hat{\beta}$, it means these samples are homogeneous.

3. Construction of the knowledge memory

In this work, many time series images and sounds are collected and converted into limited dimensional vectors of ARMA (p, q) model parameter $\hat{\alpha}$ and $\hat{\beta}$. Each time series sample has an evaluation value v.

In N images and sounds are collected, a knowledge memory containing N data points is constructed. Each data point stands for robot location. It consists of two parts, the ARMA (p,q) parameter vectors $\hat{\alpha}_i$ and $\hat{\beta}_i$, and the evaluation value v_i, i = 1, ..., N.

4. Query recognition

A *query* time series is defined as a time series observation sequence whose evaluation value is unknown. The objective of query recognition is to approximate a query time series' evaluation value.

This task can be done in three steps,

a. Estimating the parameters of the query series, $\hat{\alpha}_q$ and $\hat{\beta}_q$.

b. Calculating the distance from the query series to every data point in the memory. Let's define the distance from the query to the *k*'th data point in the memory as,

$$\mathrm{dist}^2(k) = \sum_{i=1}^{p} u_i(\hat{\alpha}_i(k) - \hat{\alpha}_i(q))^2 + \sum_{j=1}^{q} w_j(\hat{\beta}_j(k) - \hat{\beta}_j(q))^2 \quad (2)$$

in which k = 1, ..., N.

Weights, u_i and w_j, can be inserted into the distance definition and assign high weight values to those more significant parameters. If dist(k) is small enough, we claim that the query series is similar to a time series observed previously, which is represented by the *k*'th data point in this memory.

IV. EXPERIMENTS

The experimental setup includes collection video from the movement of robot along the cell is shown in Figure 1. Only 15 frames are shown due to space availability. The details of the camera are as follows:

Canon PowerShot A1100 IS

focul length=6.2mm

image size =640 X 480

Lens 6.2 -24.8mm

Aperture value 8

Shutter speed 1/60

The position of the mobile office robot is estimated based on the images and sounds already stored in the robot memory. Images are segmented using standard segmentation algorithms of the matlab software and the important features are extracted. Similarly, cepstrum values are extracted from the sounds using matlab software and they are stored in the memory. Hence features from images and sounds are stored in the memory. Time series model is applied using matlab software and the next position of the robot is estimated.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Fig. 1: Frames of a corridor along the path of the office mobile robot

VI CONCLUSION

Time series approach is used for estimating the next position of the robot. The input parameters for the model are features extracted from the images and sounds using matlab software. The next position of the robot is based on the current position of the robot corresponding to equivalent image and sound feature stored in the memory

REFERENCES

[1] Atiya S., and Hager G., 1993, Real-time vision-based robot localization, IEEE Transactions on Robotics and Automation, Vol.9, Issue 6, pp.785-800.

[2] Branca A., Stella E., Distante A., 1997, Mobile robot navigation using egomotion estimates, Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, Vol.2, pp.533-537.

[3] Briggs A., Li Y., Scharstein D., Wilder M., 2006, Robot Navigation Using 1D Panoramic Images. In: IEEE International Conference on Robotics and Automation, pp.2679–2685.

[4] Cambell J., Sukthankar R., and Nourbakhsh I., 2004, Techniques for Evaluating Optical Flow for Visual Odometry in Extreme Terrain , Proceedings of IEEE/RSJ International Conference on Intelligence Robots and Systems, Vol.4, pp.3704-3711.

[5] Cambell J., Sukthankar R., Nourbakhsh I., and Pahwa A., 2005, A Robust Visual Odometry and Precipice Detection System Using Consumer-grade Monocular Vision, Proceedings of IEEE International Conference on Robotics and Automation, pp.3412-3427.

[6] Cao Zuo L., Oh Sung J., and Hall Earnest L., 1986, Dynamic omnidirectional vision for mobile robots, Journal of Robotic Systems, Vol.3, No.1, pp.5-17.

[7] Ching-Chih T., Chin-Cheng C., Cheng-Kain C, Yi Yu L., 2010, Behavior-based navigation using heuristic fuzzy kohonen clustering network for mobile service robots, International Journal of Fuzzy Systems, Vol.12, No.1, March, pp.25-32.

[8] Emanuele Menegatti, Takeshi Maeda and Hiroshi Ishiguro Elsevier Science, 2004, Image-based memory for robot navigation using properties of omni directional images, Robotics and Autonomous Systems, Vol.47, Issue 4, 31, pp.251–267.

[9] Jwu-Sheng Hu, Chen-Yu Chan, Cheng-Kang Wang and Chieh-Chih Wang, 2011, Simultaneous Localization of Mobile Robot and Multiple Sound Sources Using Microphone Array, Advanced Robotics, Vol.25, pp.135–152.

[10] Kawamura K., Peters R.A., Johnson C., Nilas P., and Thongchai S., Supervisory control of mobile robots using sensory Egosphere, Proceedings of 2001 IEEE International Symposium on computational intelligence in robotics and automation, July 29–August 1, 2001, Canada.

[11] Krishna K.M., and Kalra P.M., 2001, Perception and remembrance of the environment during real-time navigation of a mobile robot. Robotics and Autonomous Systems, Vol.37, Issue 1, pp.25-51.

[12] Krotkov E., 1989, Mobile robot localization using a single image, Proc. IEEE Int. Conf. Robotics Automation., Vol.2, pp.978 -983.

[13] Leonard J.J., Durrant-Whyte H.F., and Cox I.J., 1992, Dynamic map building for an autonomous mobile robot, Int. J. Robotics Res., Vol.11, No.4, pp.286 -298.

[14] Nobuhiro Aihara, Hidehiko Iwasa, Naokazu Yokoya, and Haruo Takemura, 2003, Memory-Based Self-Localization Using Omni directional Images, Systems and Computers in Japan, Vol.34, Issue 5, pp.56–68.

[15] Palis Ratanaswasd, Will Dodd,, Kazuhiko Kawamura, and David C. Noelle,, 2005, Modular Behavior Control for a Cognitive Robot,

[16] Stoffler N.O., and Schnepf Z., 1998, An MPEG-Processor-Based Robot Vision System for Real-Time Detection of Moving Objects by a Moving Observer, Proceedings of the 14th International Conference on Pattern Recognition, Vol.1, pp.477-481.

[17] Stoffler N.O., Burkert T., and Farber G., 2000, Real-Time Obstacle Avoidance Using an MPEG-Processor-Based Optic Flow Sensor, 15th International Conference on Pattern Recognition, Vol.4, pp.161-166.

[18] Teruko YATA, Akihisa OHYA, Shin'ichi YUTA, 2000, Using one bit wave memory for mobile robots' new sonar-ring sensors, 2000 IEEE International Conference on Systems, Man and Cybernetics

[19] Will Dodd and Ridelto Gutierrez, 2005, The Role of Episodic Memory and Emotion in a Cognitive Robot, IEEE International Workshop on Robots and Human Interactive Communication, 13-15 Aug. 2005, pp.692-697.

[20] Yagi Y., Kawato S., and Tsuji S., 1994 , Real-time omnidirectional image sensor (COPIS) for vision-guided navigation, IEEE Trans. Robotics Automation, Vol.10, No.1, pp.11-22.

[21] Yagi Y., Nishizawa, Y., Yachida M., 1995, Map-based navigation for a mobile robot with omnidirectional image sensor COPIS. IEEE Transaction on Robotics and Automation, Vol.11, Issue 5, pp.634–648.

[22] Yoshio Matsumoto, Masayuki Inaba, Hirochika Inoue, 2003, View-based navigation using an omni view sequence in a corridor environment, Machine Vision and Applications, Vol.14, Issue 2, pp.121–128.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

AUTHORS BIOGRAPHY

Dr. C. Vijayalakshmi is currently working as Professor in Mathematics Department, SAS, VIT University, Chennai, Tamilnadu. She has more than 17 years of teaching experience at Graduate and Post Graduate level.

She has published more than Twenty research papers in International and National Journals and authored three books for Engineering colleges, Anna University. She has received Bharat Jyothi Award for outstanding services, achievements and contributions, Best Teachers award for academic excellence. Her area of specialization is Stochastic Processes and their applications. Her other research interests include Optimization Techniques, Data mining and Bio-Informatics.

Mrs.S.Sivagnana Sundari is currently working as Professor in Mathematics Department, D.B.Jain College, Thuraipakkam, Chennai - 97, Tamilnadu. She has 10 years of teaching experience. She has research papers to her credit.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Suggest an Aspect-Oriented Design Approach for UML Communication Diagram

Mohammed F.Nather
Software Engineering Dept.
University of Mosul
Mosul, Iraq

Dr.Nada N.Saleem
Software Engineering Dept.
University of Mosul
Mosul, Iraq

***Abstract*- More and more works are done on the design of the Unified Modeling Language (UML) which is designed to help us for modeling effective object oriented software , Existing Object-Oriented design methods are not mature enough to capture non-functional requirement such as concurrency, fault tolerance , distribution and persistence of a software approach. Our approach proposed to use aspect-oriented software development (AOSD) mechanisms to solve the issues for interactions of the communication diagram in UML that support only the Object-Oriented mechanisms ,thus AOSD allow to design programs that are out of reach of strict Object-Orientation and could possibly improve the structures and implementations.**

Keywords*-Aspect-Oriented Programming (AOP), Unified Modeling Language (UML) Aspect Oriented Software Development (AOSD), Software Engineering (SE), Separation of Concerns (SoC)

I. INTRODUCTION

Aspect-Oriented Software Development (ASOD) has arisen as an approach that supports a better Separation of Concerns (SoC) and more adequately reflects the way developers think about the system [3]. Essentially, AOSD introduces a unit of modular implementation – the aspect – which has been typically used to encapsulate crosscutting concerns in software systems (i.e., concerns that are spread across or tangled with other concerns). Modularity, maintainability, and facility to write software can be achieved with AOP [9]. At the early stages of software development, however, concerns strongly overlap with broadly scoped qualities such as performance, reliability, and security with relationships that are often of a qualitative nature. These relationships go well beyond common composition rules such as before, after, around, concurrent, and interleaved . For example, a security concern surely impacts negatively a performance concern, because more resources are required for security features such as encryption, authentication, or access control, but just how much on the other hand, a performance concern can affect negatively a security concern if the performance concern caches results, which must then be protected. These are examples of qualitative interactions. Interactions are one of the most interesting kinds of relationships between concerns, because they describe potentially undesirable impact of one concern on another. Interactions have been studied extensively in the telecommunications domain but are applicable to many other domains and applications. Interactions manifest themselves in aspect-oriented models, when multiple aspects affect the same elements in the base [11].

II. ASPECT-ORIENTED PROGRAMMING

AOP is merging discipline in Software Engineering, aimed at modularizing crosscutting concerns by encapsulating replicated, scattered, and tangled code in Aspects , Aspect-Oriented Software development (ASOD) is a new technique to support separation of concerns in software development [1], Examples of crosscutting concerns include tracing, logging, caching, resource pooling and so on. The techniques of AOSD make it possible to modularize crosscutting aspects of a system. Like objects in object-oriented software development, aspects in AOSD may arise at a stage of the software life cycle, including , requirements, specification, design, implementation, etc [4] .

A. Aspectj

AOP has been implemented in different languages, among them Smalltalk and Java . The Java implementation of AOP is called AspectJ (TM) and has been created at Xerox PARC [14]. AspectJ is an aspect-oriented extension to Java. The language is fully compatible with pure Java. However, it introduces new kinds of structures and new keywords to write aspects [6], which adds AOP capabilities to Java.

AspectJ includes following attributes:

- Aspect is a module for handling crosscutting concerns .Aspects are defined in terms of pointcuts, advice, and introduction .Aspects are reusable and inheritable .
- Joinpoint is a Point of execution of java program. it further includes constructor call , method call , exception handler execution .
- pointcut is a predicate that matches join points.
- advice is code that is executed at a pointcut ,which is an action taken by an aspect at a particular joinpoint,

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

different types of advice include ″around″, ″before″ and "after" advice.

- Weaving linking aspects with other application type or objects to create an advised object,this can be done at compile time (using the AspectJ compiler, for exam ple), load time, or at runtime.
- Target object is an object being advised by one or more aspects also referred to as the advised object [13].

B . ASPECT-ORIENTED MODELING AND DESIGN

AOM aims at supporting separation of crosscutting concerns at the modeling level, with the purpose of enhancing productivity, quality and reusability through the encapsulation of requirements that cut across software components. Aspect-oriented modeling (AOM) is therefore of great interest , which involves identifying, analyzing, managing, and representing crosscutting concerns. Zooming into design-level approaches Aspect-Oriented Modeling (AOM) provides support for separating concerns at the design level and has the potential to effectively tackle the complexity of developing software that deals with interdependent concerns[16]. Aspect-oriented design has the same objectives as any software design activity, i.e. characterizing and specifying the behavior and structure of the software system. Its unique contribution to software design lies in the fact that concerns that are necessarily scattered and tangled in more traditional approaches can be modularized, even with proper education understanding crosscutting concerns can be difficult without proper support for visualizing both static structure and the dynamic flow of a program. So the languages which implement AOP must have the facility to support the visualizing of crosscutting concerns[15] , this section describes our motivation to support aspects in the design level and describe them in XMI. Then, we outline the underlying technologies, UML, XML.

C. Benefits of Capturing Aspects in the Design Phase

Aspects can be identified at the requirement, design and implementation phases, though the inter-component tangling tends to occur at the implementation/coding phase. When aspects are identified, or emergent, at the implementation phase, developers often add or change aspects manually and maintain them in the source code level. Few methods have been proposed for expressing aspects in the design level. Supporting aspects at the design phase streamlines the process of aspect-oriented development by facilitating:

- Documentation and Learning: Supporting an aspect as a design construct allows developers to recognize it in the upper level of abstraction at the earlier stage of development process. Aspect designers and people learning aspect-oriented software can learn and document aspect models in more intuitive way. For example, they can visualize aspect models using a CASE tool that supports visual modeling [15][8].
- Reuse of aspect: The ease of documentation and learning leverages the reuse of aspect information, how an aspect is designed and how it is intended to affect to classes. It's easy to imagine more sophisticated ways of using aspects, such as aspect-aware CASE tools, hyperlinked documents and pattern catalogues that collects well-known and feasible aspects. These would increase the reusability of aspect-based design [15],[8].

III. UNIFIED MODELING LANGUAGE (UML)

The Unified Modeling Language (UML) is a general-purpose graphical object-oriented modeling language that is designed to visualize, specify, construct and document software systems in both structural and behavioral aspects. UML is intended to be a common way of capturing and expressing relationships and behaviors in a notation that's easy to learn and efficient to write [10]. UML Is a design language most accepted in software engineering, and is considered as a standard. It includes many useful ideas and concepts that have their roots in various individual methods and theories. UML provides numerous modeling techniques, including several types of diagrams, model elements, notation and guidelines. These techniques can be used in various ways to model different characteristics of a software system [4], UML 2 describes 13 official diagram types which classified into structural and behavioral diagrams, see figure 1 [2].

A. interaction diagrams

Classified into four types of diagrams (sequence diagram, communication diagrams, interaction overview diagrams, timing diagrams) helps you accurately model how the parts that make up your system interact [7].

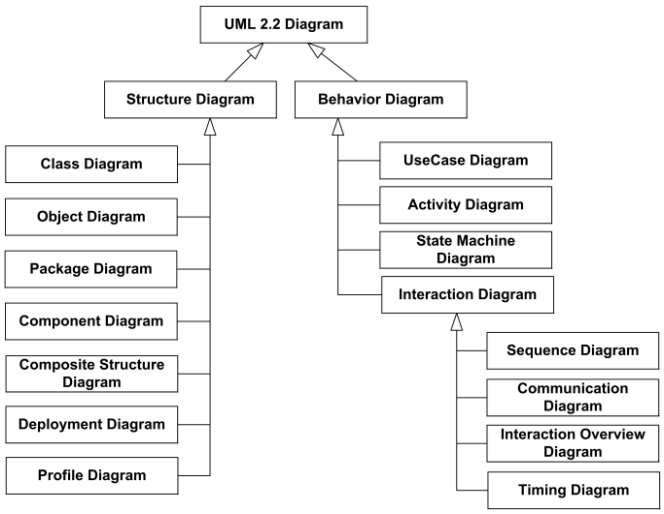


Figure 1: 1UML 2.0 Diagrams

B. Communication Diagram

A kind of interaction diagram, emphasize the data links between the various participants in the interaction. Instead of drawing each participant as a lifeline and showing the sequence of messages by vertical direction as the sequence diagrams does, the communication diagram allows free placement of participants, allows you to draw links to show how the participants connect, and use numbering to show the sequence of messages. In UML 1 .x, these diagrams were

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

called collaboration diagrams. This name stuck well, and I suspect that it will [r10].

IV. XML METADATA INTERCHANGE (XMI)

XMI is a standard that enables you to express your objects using Extensible Markup Language (XML), XMI provides a standard representation of objects in XML, enabling the effective exchange of objects using XML. It is intended to provide a standard way for programmers and other users to exchange information about metadata (essentially, information about what a set of data consists of and how it is organized). Specifically, XMI is intended to help programmers using the Unified Modeling Language (UML) with different languages and development tools to exchange their data models with each other [5].

V. OUR WORK

in this section our approach explained by the constructed software engineering tool, which produced to extend UML to support designing the communication diagram with the concepts of Aspect-Oriented modeling (AOM) .The tool input is reading XMI file obtained from other tools that describe the problem with communication diagram in Object-Oriented like EA, then parsing the XMI file to get the useful information . after that the Aspect-Oriented approach analyze the information and identify the crosscutting of the non-functional concerns , finally designing the communication diagrams with the notation of aspect-oriented and generate code for the aspect/class and the other classes for the functional concerns see figure (2) for the processing .

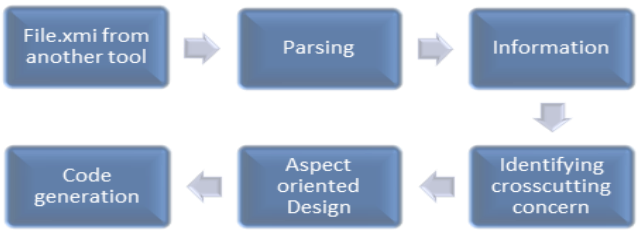


Figure 2: work processing

A. *Parsing process*

When dealing with XMI programmatically, one of the first things you have to do is take an XML document and parse it, I use Simple application program interface (API) for XML (SAX) parsing. As the document is parsed, the data becomes available to the application, see figure (3) for the XMI 1.1 generated form enterprise architect (EA) tool from sparx system products, this figure show the description of a message between two objects.

B. *Table of information*

After the parsing process the data is available for processing and saved in tables, each row contain the properties of each message between two objects, (the name of the message , object sender , object receiver , class sender , class receiver , the classes are functional or non-functional , the sequence of the message , and the diagram to which the message back) See figure (4).

C. *Identifing crosscutting and AOD*

In object-oriented programming (OOP), ones usually captures the main concerns of a problem (e.g. for a banking application, money withdrawal at an ATM, funds transfer,...) and then encapsulates them into classes. The problem is that some "secondary", non-business operations are needed (non-functional, i.e. logging, authentication, transactions), these operations are shared by several classes of a system, and each of them has to code their firing. The main idea behind AOP is to consider these operations as crosscutting aspects and to group them into separate modules, which are automatically called by the system. This new approach helps building clean software architectures and completes the toolbox of software engineers, by capturing the crosscutting depending on the number of repetition of the method call (which is the message in communication diagram) from the functional concern to non-functional concern, then compare the number with the specified threshold. The classes for the non-functional concern suggested to become an aspect/classes, then redesign each diagram that specify each operation in the system .(the problem as an input in communication diagram within OOD) , see figure(5) , and figure(6) that show how Aspect are captured and represented with red color within the notation of Aspect-Oriented design . Finally applying the fan-in and fan-out metrics to the problem in OOD to find the coupling of each class, and then apply the same metrics to the AOD.

D. *Code generation*

From figure (6) the tool can generate code for each aspect/class or other classes by double clicking on each one, see figure (7).

VI. CONCLUSION AND FUTURE WORK

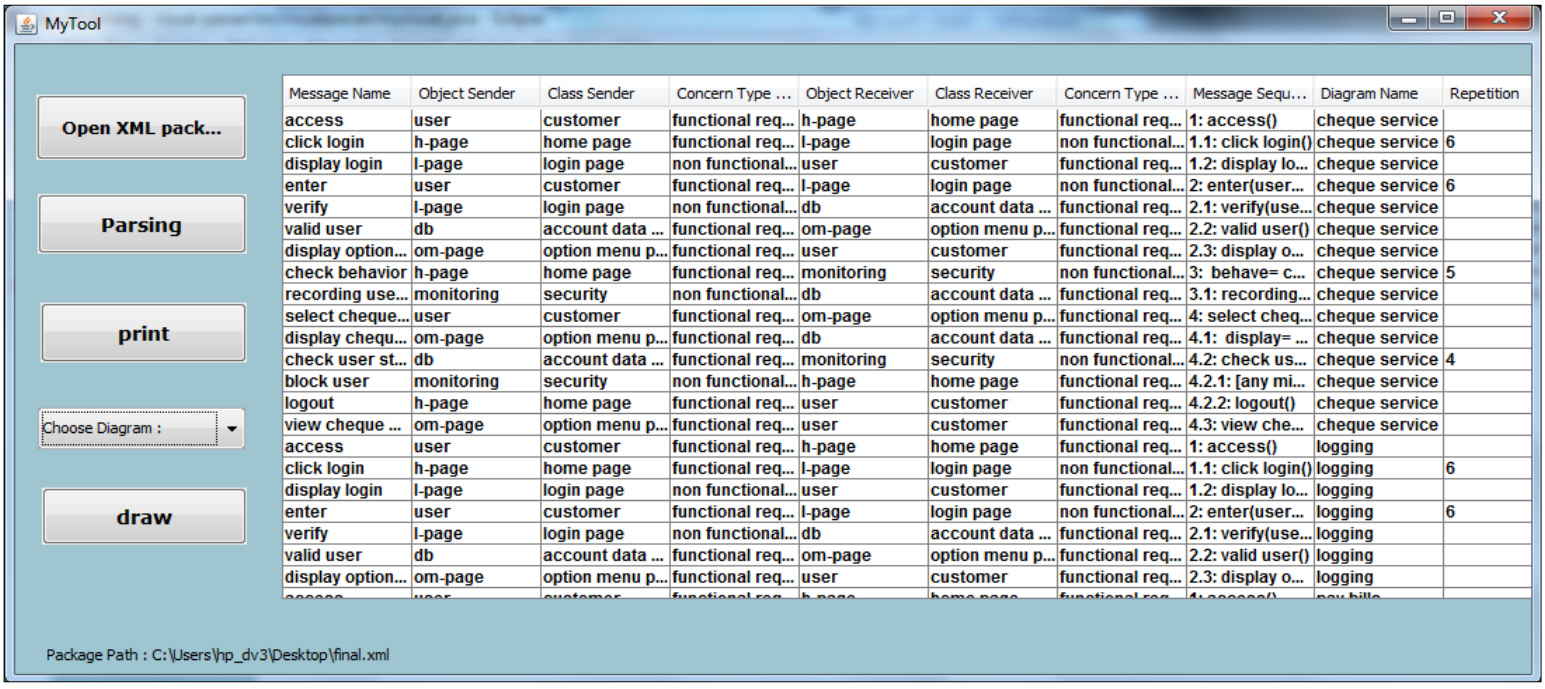
We have seen in this paper how AOP addresses non-functional requirements (crosscutting concern) using java through the constructed tool to support the work. The aim for having these crosscutting concerns being implemented separately from the core concern to bring clearly the specific implementation of our aspects to maximum understandability, which leads to better maintenance, and reusability. By applying fan-in and fan-out metrics on the problem in OOP and after redesigning the problem with AOP we found the coupling in decreased between classes and the structure is

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

improved. For future work we suggest to use artificial intelligence algorithms to identify and capture the crosscutting concern.

```
<UML:Message name="block user" xmi.id="EAID_B479EEB2_DC42_43fe_BC24_6FAA8A70B5E8" visibility="public"
sender="EAID_0DF2BC6D_B53B_4536_BC6D_2E5B175A905C" receiver="EAID_4E2A843C_3F57_4b79_8465_C696EF958116"
collaboration="EAID_1CE1F0B8_5AAD_4303_A642_3AA6AFCC51EC">
    <UML:ModelElement.taggedValue>
        <UML:TaggedValue tag="message_link" value="EAID_1CE1F0B8_5AAD_4303_A642_3AA6AFCC51EC"/>
        <UML:TaggedValue tag="style" value="1"/>
        <UML:TaggedValue tag="ea_type" value="Collaboration"/>
        <UML:TaggedValue tag="direction" value="Source -&gt; Destination"/>
        <UML:TaggedValue tag="linemode" value="1"/>
        <UML:TaggedValue tag="linecolor" value="-1"/>
        <UML:TaggedValue tag="linewidth" value="0"/>
        <UML:TaggedValue tag="seqno" value="15"/>
        <UML:TaggedValue tag="headStyle" value="0"/>
        <UML:TaggedValue tag="lineStyle" value="1"/>
        <UML:TaggedValue tag="conditional" value="any missuse"/>
        <UML:TaggedValue tag="privatedata1" value="Synchronous"/>
        <UML:TaggedValue tag="privatedata3" value="Call"/>
        <UML:TaggedValue tag="privatedata4" value="5.2.1"/>
        <UML:TaggedValue tag="ea_localid" value="190"/>
        <UML:TaggedValue tag="ea_sourceName" value="monitoring "/>
        <UML:TaggedValue tag="ea_targetName" value="h-page"/>
        <UML:TaggedValue tag="ea_sourceType" value="Object"/>
        <UML:TaggedValue tag="ea_targetType" value="Object"/>
        <UML:TaggedValue tag="ea_sourceID" value="63"/>
        <UML:TaggedValue tag="ea_targetID" value="51"/>
        <UML:TaggedValue tag="src_visibility" value="Public"/>
        <UML:TaggedValue tag="src_isOrdered" value="false"/>
        <UML:TaggedValue tag="src_targetScope" value="instance"/>
        <UML:TaggedValue tag="src_changeable" value="none"/>
        <UML:TaggedValue tag="src_isNavigable" value="false"/>
        <UML:TaggedValue tag="src_containment" value="Unspecified"/>
        <UML:TaggedValue tag="dst_visibility" value="Public"/>
        <UML:TaggedValue tag="dst_aggregation" value="0"/>
        <UML:TaggedValue tag="dst_isOrdered" value="false"/>
        <UML:TaggedValue tag="dst_targetScope" value="instance"/>
        <UML:TaggedValue tag="dst_changeable" value="none"/>
        <UML:TaggedValue tag="dst_isNavigable" value="true"/>
        <UML:TaggedValue tag="dst_containment" value="Unspecified"/>
        <UML:TaggedValue tag="diagram" value="EAID_62A38B1B_58D5_4985_A568_43A9024B9734"/>
        <UML:TaggedValue tag="lt" value="5.2.1: [any missuse]:block user()"/>
    </UML:ModelElement.taggedValue>
```

Figure 3: sample of XMI



Message Name	Object Sender	Class Sender	Concern Type ...	Object Receiver	Class Receiver	Concern Type ...	Message Sequ...	Diagram Name	Repetition
access	user	customer	functional req...	h-page	home page	functional req...	1: access()	cheque service	
click login	h-page	home page	functional req...	l-page	login page	non functional...	1.1: click login()	cheque service	6
display login	l-page	login page	non functional...	user	customer	functional req...	1.2: display lo...	cheque service	
enter	user	customer	functional req...	l-page	login page	non functional...	2: enter(user...	cheque service	6
verify	l-page	login page	non functional...	db	account data ...	functional req...	2.1: verify(use...	cheque service	
valid user	db	account data ...	functional req...	om-page	option menu p...	functional req...	2.2: valid user()	cheque service	
display option...	om-page	option menu p...	functional req...	user	customer	functional req...	2.3: display o...	cheque service	
check behavior	h-page	home page	functional req...	monitoring	security	non functional...	3: behave= c...	cheque service	5
recording use...	monitoring	security	non functional...	db	account data ...	functional req...	3.1: recording...	cheque service	
select cheque...	user	customer	functional req...	om-page	option menu p...	functional req...	4: select cheq...	cheque service	
display chequ...	om-page	option menu p...	functional req...	db	account data ...	functional req...	4.1: display= ...	cheque service	
check user st...	db	account data ...	functional req...	monitoring	security	non functional...	4.2: check us...	cheque service	4
block user	monitoring	security	non functional...	h-page	home page	functional req...	4.2.1: [any mi...	cheque service	
logout	h-page	home page	functional req...	user	customer	functional req...	4.2.2: logout()	cheque service	
view cheque ...	om-page	option menu p...	functional req...	user	customer	functional req...	4.3: view che...	cheque service	
access	user	customer	functional req...	h-page	home page	functional req...	1: access()	logging	
click login	h-page	home page	functional req...	l-page	login page	non functional...	1.1: click login()	logging	6
display login	l-page	login page	non functional...	user	customer	functional req...	1.2: display lo...	logging	
enter	user	customer	functional req...	l-page	login page	non functional...	2: enter(user...	logging	6
verify	l-page	login page	non functional...	db	account data ...	functional req...	2.1: verify(use...	logging	
valid user	db	account data ...	functional req...	om-page	option menu p...	functional req...	2.2: valid user()	logging	
display option...	om-page	option menu p...	functional req...	user	customer	functional req...	2.3: display o...	logging	

Figure 4: Table showing the information after parsing

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

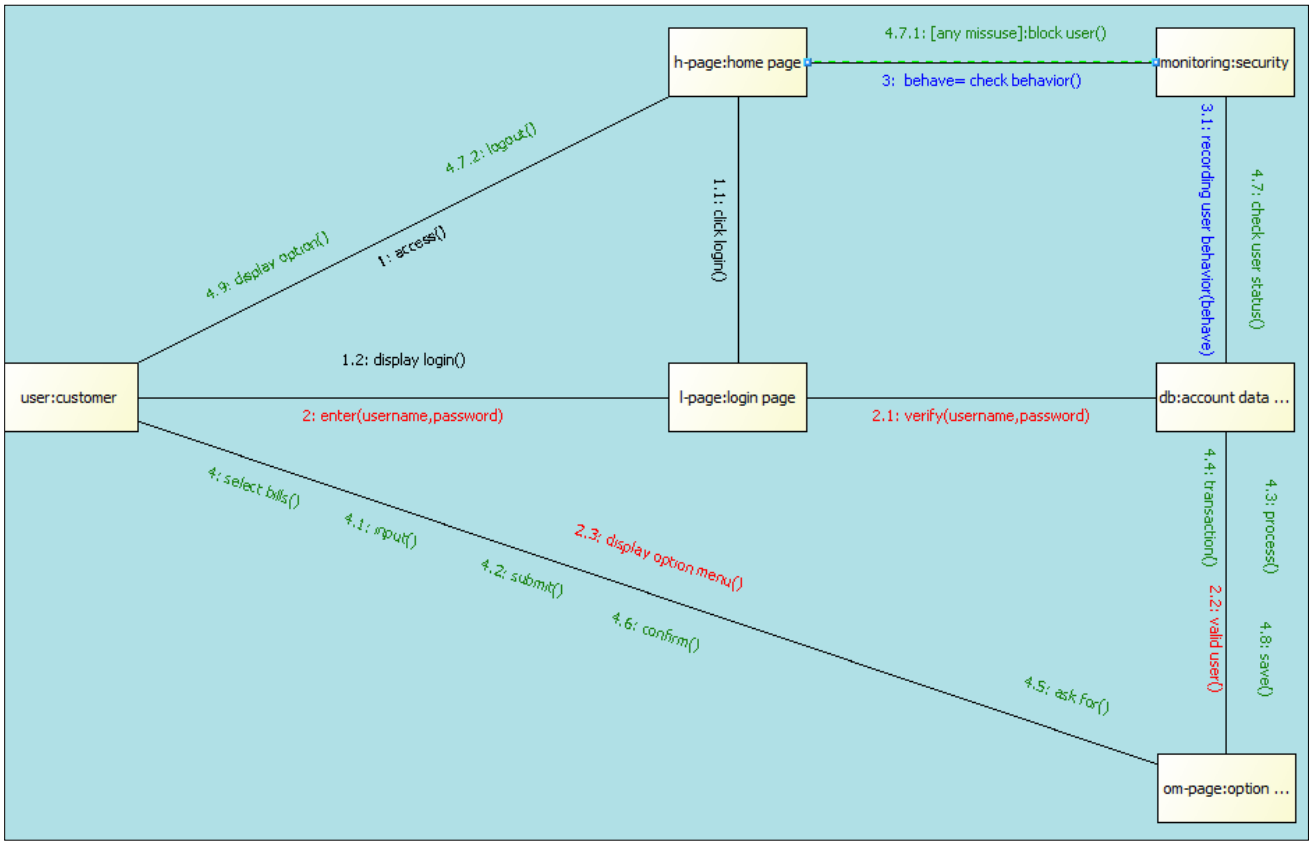


Figure 5: communication diagram in OOP

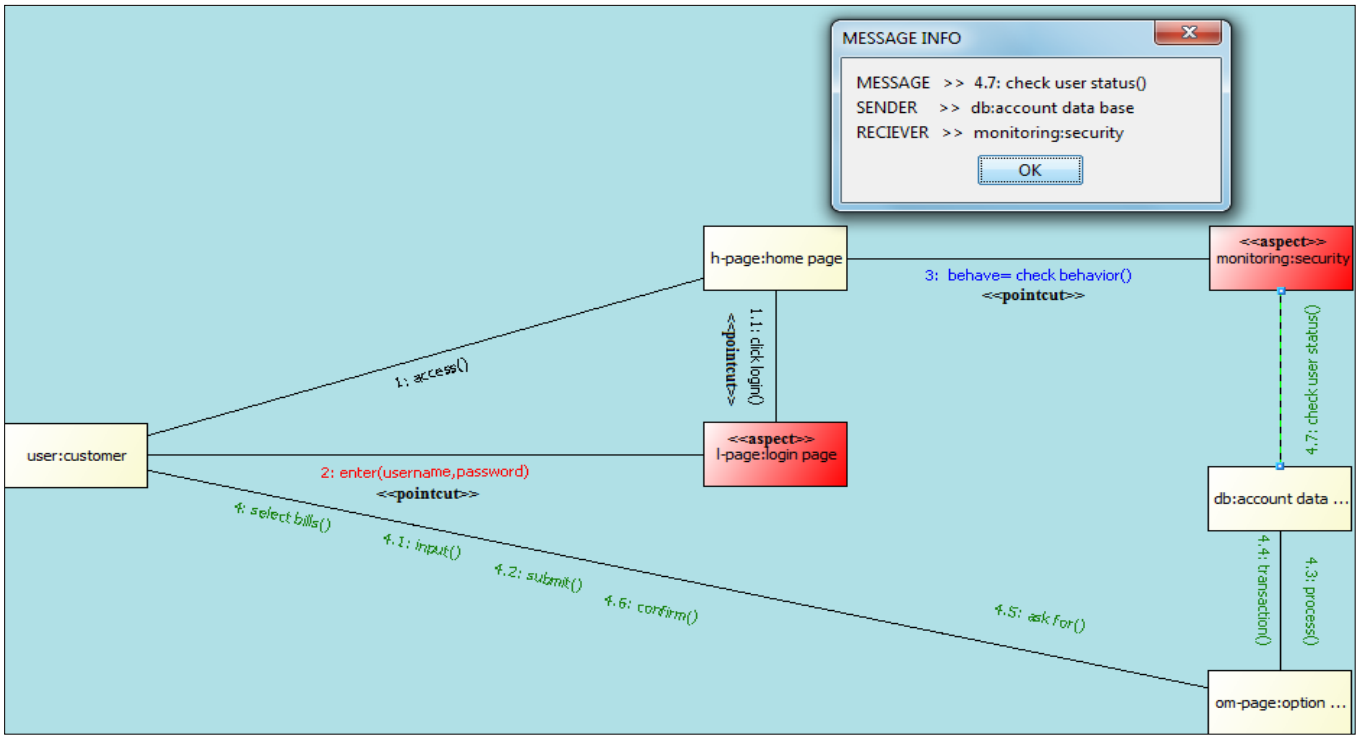


Figure 6: Applying AOD approach

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

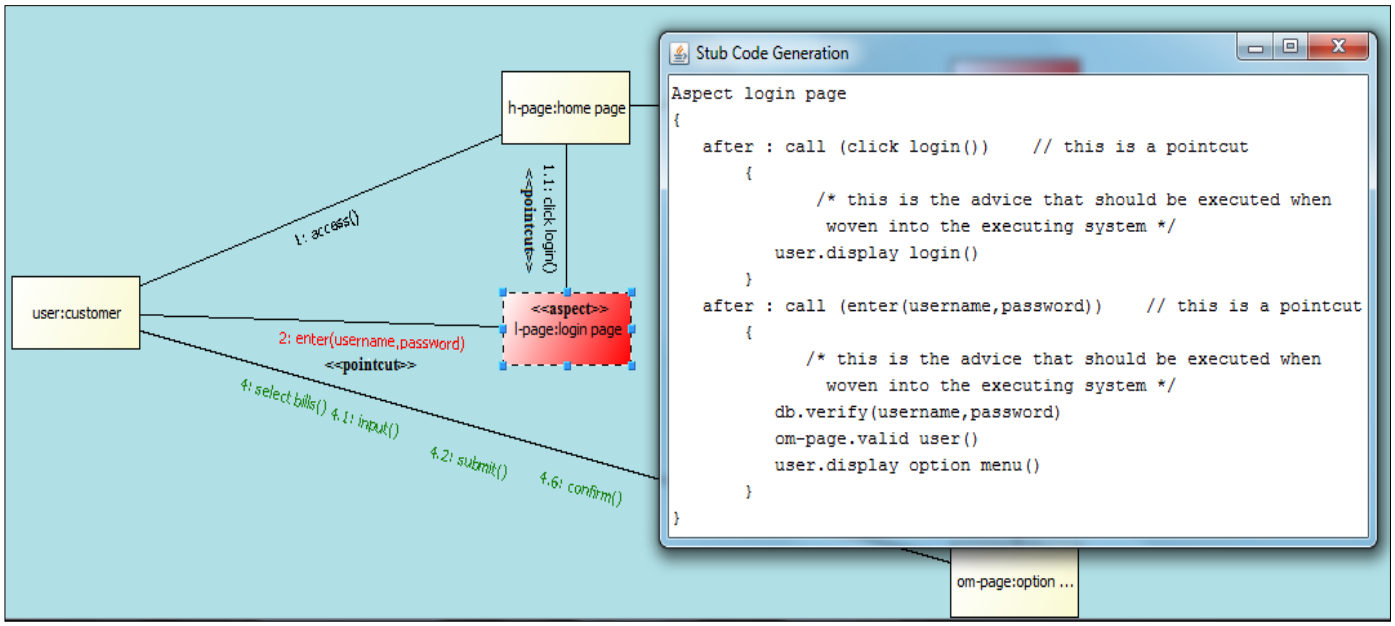


Figure7: Code generation from AOD

I. REFERENCES

[1] B.lodewijk and Mehmet, concerns using composition filters, communications of the acm, volume 44, no. 10, 2001.

[2] F.martin, uml distilled, a brief guide to the standard object modeling language, third edition, Addison wisely,2004.

[3] G.kiczales, j. irwin, j. lamping, j. loingtier, C. lopes, C. maeda, and A. menhdhekar, "aspect-oriented programming," in proceedings of the 11 the European conference on object-oriented programming (ecoop), jyväskylä, Finland, 1997, pp. 220–242, 1997.

[4] G .sushil , k. s. kahlon and P. k. bansal ,how to measure coupling in aop from uml diagram ,international journal of computer science and telecommunications ,volme2,issue 8,2011.

[5] G.timothy, d.gracy, b.stephen, mastering xmi: java programming with xmi, xml, and uml, Wiley computer publishing,2002

[6] H.bruno, separating concerns in scientific software using aspect-oriented programming, 2006, thesis at the University of Manchester for the degree of doctor of philosophy in the faculty of engineering and physical sciences

[7] H. kim, M. russell, learning uml 2.0, o'reilly, 2006.

[8] k. czarnecki. generative programming: principles and techniques of software engineering based on automated configuration and fragment-based component models, dissertation in Department of Computer Science and Automation, Technical University of Ilmenau, 1998.

[9] M.ana, G. john (editors), early aspects: current challenges and future directions, springer, lncs 4765, pp 19-38, 2007.

[10] M.elkamel, S.halima, M.nabil and C.allaoua, design of atl rules for transforming uml 2 communication diagrams into buchi automata, international journal of software engineering and application, vol7,no 2,2013.

[11] R. laddad, "aspect-oriented programming will improve quality," IEEE, vol. 20, no. 6, pp. 90–91, 2003.

[12] S. junichi ,Y.yoshikazu ,extending uml with aspects: aspect support in the design phase,submitted to the 3rd aspect-oriented programming (aop) workshop at ecoop'99, 1999

[13] S.kotrappa , K. prakash, stronger enforcement of security using aop & spring aop ,journal of computing, volume 2, issue 6,2010

[14]S. halse and S. patil , paper on aspect-oriented programming with aspectj programming approach,j. comp. & math. sci. vol.2 (4), 637-646, ,2011.

[15] Wikipedia, the free encyclopedia, aspect-oriented modeling and design, 2009.

[16] Z.jing ,C. thomas ,B. aswin , and g.jeff , aspect composition in the motorola aspect-oriented modeling weaver,journal of object technology,vol. 6, no. 7,2007

SATELLITE IMAGE CLASSIFICATION METHODS and LANDSAT 5TM BANDS

Jamshid Tamouk
Department of Computer Engineering
EMU University
Famagusta, North Cyprus

Nasser Lotfi
Department of Computer Engineering
EMU University
Famagusta, North Cyprus

Mina Farmanbar
Department of Computer Engineering
EMU University
Famagusta, North Cyprus

***Abstract*—This paper attempts to find the most accurate classification method among parallelepiped, minimum distance and chain methods. Moreover, this study also challenges to find the suitable combination of bands, which can lead to better results in case combinations of bands occur. After comparing these three methods, the chain method over perform the other methods with 79% overall accuracy. Hence, it is more accurate than minimum distance with 67% and parallelepiped with 65%. On the other hand, based on bands features, and also by combining several researchers' findings, a table was created which includes the main objects on the land and the suitable combination of the bands for accurately detecting of landcover objects. During this process, it was observed that band 4 (out of 7 bands of Landsat 5TM) is the band, which can be used for increasing the accuracy of the combined bands in detecting objects on the land.**

Keywords: parallelepiped, minimum distance, chain method, classification, Landsat 5TM, satellite band

I. INTRODUCTION

The focus of this paper is solely on some satellite image classification methods for land and also Landsat 5TM bands and the suitable combination of them to have higher accurate results during classification.

A. Classification methods

The basic principle of the classification is classifying of images based on placing pixels with similar brightness value into the same group. It is done by selecting limited area or instance from images and then to assign the label (i.e. name) and color to that area. The images of the satellite (Landsat 5TM) which are used in this paper, has 7 bands for capturing the image of the earth. Each of these bands uses different wavelength for capturing the images, in a way that it causes to have 7 images from the same area but with different characteristics. The Landsat 5TM bands descriptions are shown in the Table 1:

TABLE 1: DESCRIPTION OF BANDS IN LANDSAT 5 TM [11]

Band	Wavelength(μm)	Spectral	Resolution (m)
1	0.45 – 0.52	Blue-Green	30
2	0.52 – 0.60	Green	30
3	0.63 – 0.69	Red	30
4	0.76 – 0.90	Near IR	30
5	1.55 – 1.75	Mid-IR	30
6	10.40 – 12.50	Thermal IR	120
7	2.08 – 2.35	Mid-IR	30

Methods of classification mainly follow two approaches, namely supervised and unsupervised classification. Supervised classification is the classification that needs to interact with user (i.e. training the system) who has knowledge (ground truth) about that area before image processing. However, in unsupervised classification, it is not necessary to have high knowledge about areas and it does not need to train the system. System starts grouping the pixels, which are similar in brightness value into unique clusters. Then after finishing clustering, the user will start to label each of the groups (classes).

The following two tables, which are resulted from some of the classification methods' comparison introduce by different researchers. According Table 2 (Hosseini et. al), overall accuracy of minimum distance (73.77%) is much better than parallelepiped method (34.27%) maximum likelihood method (with 85.83% overall accuracy) provides a higher accuracy than minimum distance.

TABLE 2: ACCURACY OF DIFFERENT CLASSIFICATION METHODS AND ALGORITHMS [6]

Supervised Classification Method	Overall Accuracy %
Paralleleped	34.27
Minimum Distance	73.77
Maximum Likelihood	85.83

Table 3 is the part of comparison table of classification methods, which is given by Todd [12]. According to this table, maximum likelihood method with 90.2% accuracy is

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

a more accurate supervised classification in comparing with minimum distance (with 75.5% accurate) and parallelepiped (with 87.1% accurate).

TABLE 3: ESTIMATED ACCURACY AND PROCESSING TIME OF SOME CLASSIFICATION METHODS[12]

Method	Estimated process time	Accuracy %
Maximum Likelihood	18 min	90.2
Minimum Distance	15 min	75.5
parallelepiped	15 min	87.1
ISODATA	2.25 min	90.6

In comparing accuracy of maximum likelihood (supervised classification) with ISODATA method (unsupervised classification) according to the above table, we can see that their accuracies are approximately same. Here can see that opposite the previous table (table 1.2) parallelepiped is much accurate than minimum distance.

Based on the accuracy tables for supervised classification in the above tables (researches), maximum likelihood classification method is the most accurate method comparing with parallelepiped and minimum distance to mean methods and about parallelepiped and minimum distance there are different ideas. Some of researchers' finding like Hashemi el. al [4], Oruc el. al [9] and Todd [12] shows parallelepiped is more accurate than minimum distance but the others findings like Lim el. al [7] and Hosseini el. al [6] show the opposite of that.

B. 1.2 Suitable combination of the bands for land covers

The other issue is about the way of combining the different bands of satellite for achieving a good result. In other words, for having more accurate result from classification of satellite images we should consider the selection of the suitable or correct combination of bands (according to the object or objects on the land, which we want to classify). According to the following formula, "[n! / (r! (n-r)!)]", 7 bands with three-band combination can have, 35 kinds of combination of bands. Below is the table of the combination of different bands and the corresponding "OIF Index" values, where OIF is used to show how useful is the combination of the bands, based on the correlations of them:

Table 4 below, which made by Wenbo el. al [13] for 20 selected bands combinations in to ascending order. In this table we can find that, OIF index of the combination of bands TM 3, 4, 5 (ETM+3, 4, 5) is the highest OIF index.

In this table from best 10 three-band combinations out of 20 three-band combinations, 8 of them include band 4 which gives highest number of occurrence in the combinations. It means 80% of 10 best three-band combinations (13**4**, 157, 357, 2**4**5, 1**4**5, 2**4**7, 1**4**7, **4**57, 3**4**7, and 3**4**5) have band **4** in their combination.

TABLE 4: OIF INDEX OF DIFFERENT BANDS COMBINATION BY ASCENDING ORDER [13]

Band combination	IOF index	Range Order	Band combination	IOF index	Range Order
123	12.832	1	13**4**	22.605	11
127	16.739	2	157	22.724	12
124	17.229	3	357	22.840	13
237	18.043	4	2**4**5	23.918	14
125	18.359	5	1**4**5	24.316	15
137	19.160	6	2**4**7	24.858	16
135	19.693	7	1**4**7	25.724	17
235	20.596	8	**4**57	27.442	18
257	21.314	9	3**4**7	29.209	19
234	22.169	10	3**4**5	29.230	20

In Table 5, which made by Hobson [5] all possible combination of bands (35 set of combined bands) in three-band combination has been examined which the best combination of the bands belongs to the bands 4, 5, 6 with 57.3673 OIF. According to the below table it can be understood that from the best 10 tree-band combination out of 35, eight of these combination include band 4 in their combination. It means 80% of 10 best three-band combinations (145, 457, 167, 246, 347,146, 346, 356, 467, and 456) have band 4 in their combination.

TABLE 5: OIF OF 35 COMBINED BANDS [5]

Bans Combination	OIF	Bans Combination	OIF
123	12.6385	237	16.1890
124	18.9822	245	27.8149
125	20.2492	2**4**6	35.9016
126	15.6910	247	23.4820
127	15.0502	256	28.8717
134	22.7656	257	22.9567
135	22.8736	267	20.3396
136	18.8254	345	31.3270
137	17.5892	3**4**6	42.0967
1**4**5	31.9984	3**4**7	39.9820
1**4**6	40.1405	356	44.0859
147	26.9859	357	30.0630
156	29.5769	367	24.4388
157	24.5532	**4**56	57.3673
167	35.7702	**4**57	33.7486
234	19.5080	**4**67	46.6954
235	21.0872	567	31.8615

According to other researchers' findings mentioned below and after finding the best bands combination for classification of land covers, now we want to see which combination of bands are suitable for each of the main objects (such as water, vegetation, soil, snow and ice, sand and so on) on the land.

In order to recognize the water on land surface or distinguish between land and water (coastal and see) a combination of band2 and band5 can be used as below:

The ratio band2/band5 is greater than one for water and less than one for land in large areas of coastal zone. With this method, water and land can be separated directly.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Another method is to use the band ratio between band 4 and band 2 [1]. So, if the result of band2/band5 become greater than one (band2/band5>1) then the object will be water; otherwise, it will be land or any other objects. Also in the following combination of bands is used for calculating the water index and classifying the images based on water and non-water by Hosseini el. al [6]:
Water Index = (Band 1 + Band 2 + Band 3) / (Band 4 + Band 5 + Band 7) [4]. That is the ratio of visible spectrum bands to be reflected by infrared bands.
The other important object on the land is vegetation or the area with vegetative cover. According to the Shan long el. al [10] finding, the best combination of bands for detecting or recognizing the vegetation is combination of band4 and band3.
NDVI = (Band 4 - Band 3) / (Band 4 + Band 3) [10]
This combination of bands 3 and 4 is called "normalized difference vegetative index" (NDVI). In above formula the bands which are used are NIR = Reflectance in Near Infrared Band (band 4) and RED = Reflectance in the RED Band (band 3). It is often used in detection of small differences between vegetation classes and sometimes used for distinguishing vegetative area from other areas or objects. In these bands, vegetation and soil contrast is at its maximum. It means soil and vegetation can be easily differentiated from each other.
The NDVI (Normalized Difference Vegetative Index) values in the range of -1.0 to 1.0, where Vegetated areas will typically have values greater than zero and Negative values indicate non-vegetated surface features such as water, barren, ice, snow, or clouds [10].
So, if NDVI is greater than zero (if NDVI > 0) then that area is the vegetative cover area; otherwise, it is land (other objects). Following is the other vegetative index formulas, which are discussed by Muzein [8]:
Corrected Naturalized Differential Vegetation Index

$$\left(\frac{Band4-Band3}{Band4+Band3}\right)\times\left[1-\frac{Band5-(Band5)_{MIN}}{(Band5)_{MAX}-(Band5)_{MIN}}\right]$$

Percent Vegetation Cover:

$$(\text{Standardized}\quad \text{NVDI})^2$$

Simple Ratio $\left(\frac{Band4}{Band3}\right)$

Reduced Simple Ratio:

$$\left(\frac{Band4}{Band3}\right)\times\left[1-\frac{Band5-(Band5)_{MIN}}{(Band5)_{MAX}-(Band5)_{MIN}}\right]$$

Soil Adjusted Vegetation Index: Minimizes the secondary backscattering effect of canopy-transmitted soil background reflect radiation. It describes both vegetation cover and soil background.

$$\left(\frac{Band4-Band3}{Band4+Band3+1}\right)\times(1+L)$$

L depends on Land cover, but 0.5 is a suggested value for many land cover conditions
According to Hashemi el. al [4] research finding for discriminating or distinguishing between salt and sodium soils, use of Landsat 5TM bands (2,4,6) and 7 are more accurate than other combination of the bands.
Also according to the Hashemi el. al [4] finding the ratio spectral (TM3 - TM4) / (TM2 - TM4), show the best correlation with the Soil EC data. It means that by combination of the band2, band3 and band4 we can distinguish different kinds of soil from each other and other objects.
The other objects on surface of the land, are snow and ice (i.e. glacier area) which based on the Todd research finding [12] can be detected or recognized by using the combination of bands 4 and 5 (ice or snow index = band4 / band5), bands 3 and 5 (ice or snow index = band3 / band5), or bands 3, 4 and 5.
"Glacier extent mapping from satellite data has been the focus of many recent research papers. Bayr et al. (1994) used a threshold of a ratio image of TM 4 to TM 5 bands to delineate glacier area, while Rott (1994) used a threshold of a TM 3 to TM 5 ratio image. Paul (2000) found that the TM 4 to TM 5 ratio technique yielded the best results for glacier mapping on Gries Glacier, especially in regions with low insolation" [12].
"Also using the visible-red channel (TM 3) and two of the infrared channels (TM 4 and 5) allowed for an excellent distinction of the ice cap. This is because snow and ice have very high spectral reflectance in the visible-red (RED) and the near-infrared (NIR) wavelength regions and very low spectral reflectance's in the middle-infrared (MIR) wavelength region" [12].
Based on the above discussion, the best combination for distinction of ice and snow is combination of bands 3, 4 and 5.
Below there are some other combinations of the bands which are, the findings of several researchers. Based on the result of the below researches, combination of bands 3, 4, 7 is good for detecting the water boundary or costal, soil moisture and iron compounds. Bands 4,3, 2 are used for vegetation and crop analysis, bands 4, 5, 3 for soil moisture and vegetation analysis, bands 3, 2, 1 for landcover and underwater features, bands 7, 4, 3 and bands 7, 4 , 2 for change detection, soil type and vegetation stress. [2, 3 and 7]

II. PROBLEM DEFINITION

Researchers depending on their application should be able to choose suitable method for classification of their images. Obviously, it is difficult for the researchers who are new in this field. On the other hand, most of the researchers who use Landsat images in their research need to know more about the Landsat 5TM bands and usage of each band. In

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

addition, they should know which bands and which combination of the bands are good for detecting different kinds of objects on the land. Therefore, they will be able to get a good result if they are able to choose suitable methods and also suitable bands or combination of the bands for their research.

III. PROPOSED METHOD

By using Landsat 5TM images as the input data (captured from north of Iran) and with training the system by the information collected about that area (ground truth) for supervised methods and without training the system for unsupervised method achieved to the ability to classify satellite images and also calculate the accuracy of each methods. Then the classification methods (supervised and unsupervised classification) are compared based on the determined accuracy. On the other hand, Table 8 is created based on the bands' features to illustrate the main objects on the land and the suitable combinations of bands for recognizing them. According to this table and by checking the other research findings on bands and band combinations, it becomes possible to find the most effective band among 7 bands of Landsat 5TM.

IV. PERFORMANCE ANALYSIS

Accuracy is calculated by dividing the number of object/class's pixels correctly classified; over the total number of pixel belong to that object/class. In other word, Accuracy= number of pixels assigned to the correct class / number of pixels that actually belong to that class or object. If this calculation is done for all of objects/classes together, the result will call Overall accuracy. Table 6 shows the confusion matrix of minimum distance method, fund for this paper (It is a table that shows the correct and incorrect number of assigned objects to each class. It is used for computation of accuracy).

TABLE 6: ERROR MATRIX TABLE FOR MINIMUM DISTANCE TO MEAN

Class type form classification	Class types determined from reference source					
	#Plots	**Water**	**Forest**	**Grassland Agriculture**	**Mountain and soil**	**Totals**
	Water	39	0	8	4	52
	Forest	0	32	12	4	48
	Grassland & Agriculture	4	8	33	8	52
	Mountain	0	8	8	24	40
No ground truth pixels		43	48	61	40	192

V. RESULT AND DISCUSSION

Table 7 is the accuracy table which is created based on the parallelepiped, minimum distance and chain methods error matrix tables gained for this paper. According to this table, the chain method (with 79% accuracy) has highest accuracy in comparing with two others mentioned methods in this paper. Also it is found that minimum distance (with 67% accuracy) has higher accuracy than parallelepiped (with 65% accuracy). Some of researchers' findings about accuracy of parallelepiped and minimum distance approximately are similar to my research finding such as Table 2 but some others are exactly opposite such as Table 3. The reasons for this could be one of the following reasons: lack of enough data for training or testing, samples distribution, difficulty with selecting sufficient training data for supervised methods or the insufficient skill of the trainers.

TABLE 7: USER, PRODUCER AND OVERALL ACCURACY OF PARALLELEPIPED, MINIMUM DISTANCE AND CHAIN METHOD

Methods	Objects	Water	Forest	Agriculture	Mountain
Paralleled piped	**User accuracy**	68%	72%	50%	71%
	Producer accuracy	92%	62%	54%	50%
	Overall accuracy	[(43+31+27+23)/192]*100=**65%**			
Minimum Distance	**User accuracy**	77%	67%	62%	60%
	Producer accuracy	91%	67%	53%	60%
	Overall accuracy	[(39+32+33+24)/192]*100=**67%**			
Chain Method	**User accuracy**	86%	77%	73%	80%
	Producer accuracy	92%	67%	73%	89%
	Overall accuracy	[(45+41+32+34)/192]*100=**79%**			

Below is the table of some important objects and the suitable combination of the bands for detecting those objects based on the literature given in section B.

TABLE 8: COMBINATION OF BANDS BASED ON THE TYPE OF OBJECT

	Bands							
Objects	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Combination of the bands & Conditions**
Water		*			*			TM 2/TM 5>1
		*		*				TM 2/TM 4>1
	*	*	*	*	*		*	Index=(TM 1+TM 2+TM 3)/(TM 4+TM 5+TM 7)
		*	*	*				Water appears dark
Coastal			*	*			*	
Vegetation			*	*				(TM 4-TM 3)/(TM 4+TM 3)>0
		*	*	*				False color infrared
			*	*	*			Vegetation conditions
Snow and Ice				*	*			Index=TM 4/TM 5
			*		*			Index=TM 3/TM 5
			*	*	*			
Soil type			*	*			*	
		*		*			*	
Salt and Sodium Soil		*		*		*	*	
		*	*	*				(TM 3-TM 4)/(TM 2-TM4)
Iron Compounds			*	*			*	Such as ilmenite
Soil moisture differences			*	*			*	Best combination
			*	*	*			
Change detection, disturbed soils vegetation stress			*	*			*	
		*		*			*	

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Table 8 shows the important object on the land and the best combination of the bands for detecting them. It is important to know that which combination of the bands can detects which kind of object on the land with more accuracy. In Table 8, by carefully looking at this table, we can find that band 4 is used in all of the objects in most of the combined bands (around 90% of the combined bands). It can also be observed in Table 4 that from the best 10 three-band combinations out of 20, eight of them (i.e. 80% of ten best three-band combinations) include band 4 in their combinations (134, 157, 357, 245, 145, 247, 147, 457, 347, and 345). The same result is approximately shown in Table 5. In this table from the best 10 three-band combinations out of 35, eight of these combinations include band 4 in their combinations (145, 457, 167, 246, 347,146, 346, 356, 467, and 456) which is 80% of 10 best three-band combinations. Therefore, these findings confirm that band 4 is the most useful band to increase the accuracy of combined bands in detecting the objects on the land.

VI. CONCLUSION

According to this paper, the proposed chain method with 79% accuracy is more accurate than the other two compared methods. In addition, Table 8, which identifies the suitable band combinations for each of the main objects on the land, was created. Finally, after analyzing the findings of this paper and some other researchers, similarly it is concluded that, having band 4 (Near Infrared) in the combinations of the bands can improve the accuracy of detection and classification of the objects noticeably.

REFERENCES

[1] Alesheikh, A.A., Ghorbanali, A and Nouri, N. (2007). *Coastline change detection using remote sensing.* ISSN: 1735-1472, IRSEN, CEERS, IAU.

[2] CURRENT SCIENCE. (2008). *Discovery of heavy mineral-rich sand dunes along the Orissa–Bengal coast of India using remote sensing techniques*. CURRENT SCIENCE, VOL. 94, NO. 8, 25.

[3] Davis, A and Allen, J. *Landsat 7 at NASA Goddard Space Flight Center*. [online] viewed (December 2012) http://landsat.gsfc.nasa.gov

[4] Hashemi, S.S., Baghernejad, M., Pakparvar, M and Emadi, M. (2005). *GIS classification assessment for mapping soils by satellite images.* Soil Science Dept, College of Agriculture. Shiraz University. Shiraz, Iran.

[5] Hobson, V. R. (2003). *Characterization of Craters of the Moon Lava Field using Landsat TM data*. FinalReport FRWS 6750 – Applied Remote Sensing.

[6] Hosseini, S.Z., Khajeddin, S.J and Azarnivand, H. (2008). *Application of ETM+data for estimating rangelands cover percentage*. College of Natural Resources & Desert Studies', University of Yazd, Yazd, Iran.

[7] Lim, H.S., MatJafri, M.Z and Abdullah, K. (2002). *Evaluation of conventional digital camera scenes for thematic information extraction.* School of Physics Universiti Sains Malaysia, Penang, Malaysia

[8] Muzein, B.S. (2006). *Remote Sensing & GIS for Land Cover/ Land Use Change Detection and Analysis in the Semi-Natural Ecosystems and Agriculture Landscapes of the Central Ethiopian Rift Valley*. PHD research in Technische Universität Dresden.

[9] Oruc, M., Marangoz, A. M and Buyuksalih, G. (2004). *Comparison of pixel based and object oriented classification approaches using Landsat 7ETM spectral bands.*

[10] Shan-long, L., Xiao-hua, S and Le-jun. Z. (2006). *Land cover change in Ningbo and its surrounding area of Zhejiang Province.* Journal of Zhejiang University SCIENCE A ISSN 1009-3095 (Print); ISSN 1862-1775 www.zju.edu.cn/jzus.

[11] Short, N. M. *Technical and historical perspectives of remote sensing.* [online] viewed (December 2012), Available at: http://ltpwww.gsfc.nasa.gov/IAS/handbook/handbook_toc.html.

[12] Todd H. A. (2002). *Evaluation of remote sensing techniques for ice area classification applied to the tropical quelccaya icecap*, Peru, Polar Geography, 2002, 26, No. 3.

[13] Wenbo,W., Jing, Y and Tingjun, K. (2008). *Study of remote sensing image fusion and its application in image classification*, *The International Archives of the Photogrammetry.* Remote Sensing and Spatial Information Sciences. Vol. XXXVII.

A SURVEY OF SATELLITE IMAGERY CLASSIFICATION WITH DIFFERENT APPROACHES

Dr. Ghayda A. Al-Talib
Dept. of Computer Sciences, College of Mathematics and Computer sciences
University of Mosul
Mosul, Iraq

Ekhlas Z. Ahmed
Dept. of Computer Sciences, College of Mathematics and Computer sciences
University of Mosul
Mosul, Iraq

***Abstract*— This paper, proposes a new classification method that uses Hidden Markov Models (HMM s) to classify remote sensing imagery by exploiting the spatial and spectral information. When applying unsupervised classification to remote sensing images it can provide more useful and understandable information. Experiments shows that other clustering scheme like traditional k-means does not performs well because it does not take into account the spatial dependencies. Experiments are conducted on a set of multispectral satellite images. Proposed algorithm is verified for simulated images and applied for a selected satellite image processing in the MATLAB environment.**

***Index Terms*— Hidden Markov Models(HMM), land cover, multispectral satellite images, unsupervised classification**.

I. INTRODUCTION

In this paper the Hidden Markov Models (HMM s) for unsupervised satellite image classification has been used. HMMs were extensively and successfully used for texture modeling and segmentation[1]. Image classification refers to the computer-assisted interpretation of remotely sensed images. Mainly, there are two ways to do the remote sensing image classification. One is visual interpretation, and the other is computer automatic interpretation[2]. The classification is an important process, which made the raw image data more meaningful information. The aim of image classification is to assign each pixel of the image to a class with regard to a feature space. These features can be considered as the basic image properties like intensity, amplitude, or some more advanced abstract image descriptors as textures which can also be exploited as a feature[3]. The computer automatic classification of remotely sensed imagery has two general approaches, supervised and unsupervised classification[4]. In this work the intensity property of the satellite images was proposed to classify the land cover. The model parameters are set randomly and then will estimated the optimal values with Baum-Welch algorithm which is the most widely adopted methodology for model parameters estimation [5]. After the model parameters are well estimated the clustering and classification process will be done.

II. HIDDEN MARKOV MODELS

HMM was distinguished from a general Markov model in that the states in an HMM cannot be observed directly (i.e. hidden) and can only be estimated through a sequence of observations generated along a time series. Assume that the total number of states being N, and let q_t and o_t denote the system state and the observation at time t. HMM can be formally characterized by $\lambda=(A, B, \pi)$, where A is a matrix of probability transition between states, B is a matrix of observation probability densities relating to states, and π is a matrix of initial state probabilities, respectively. Specifically, A, B, and π are each further represented as follows[4]:

$$A=[a_{ij}],\ a_{ij} = P(q_{t+1}=j \mid q_t = i),\ 1 \le i,j \le N \quad (1)$$

Where

$$a_{ij} \ge 0, \sum_{j=1}^{N} a_{ij} = 1, \text{for } i=1,2,\ldots,N \quad (2)$$

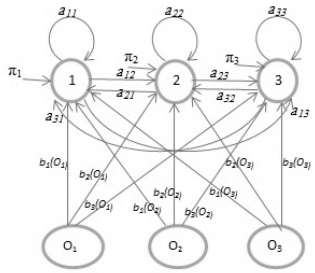


Fig.1. HMM parameters A, B and π in the case of t=1.

$$B=[b_j(o_t)],\ b_j(o_t)=P(o_t \mid q_t=j),\ 1 \le j \le N \quad (3)$$

$$\pi=[\pi_i],\ \pi_i=P(q_1=i),\ 1 \le i \le N \quad (4)$$

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

where

$$\sum_{i=1}^{N} \pi_i = 1 \quad (5)$$

For illustration purposes, HMM and related parameters *A, B*, and π are shown in Fig. 1. The observation probability density $b_j(o_t)$ for state j given observation o_t is generally modeled in Gaussian distribution as in Eq. 6:

$$b_j(o_t) = \frac{1}{(2\pi)^{k/2}|\Sigma_j|^{1/2}} \exp\left[-\frac{1}{2}(o_t - \mu_j)' \Sigma_j^{-1}(o_t - \mu_j)\right], \quad (6)$$

where (‘) prime denotes vector transpose and *k* is the dimension of observation vector o_t.

Given HMM, λ and observation sequence $O=\{o_1, o_2, \ldots, o_T\}$, one may estimate the best state sequence $Q^*=\{q_1, q_2, \ldots, q_T\}$ based on a dynamic programming approach so as to maximize $P(Q^*|O, l)$, [3]. In order to make Q^* meaningful, one has to well set up the model parameters *A, B* and π. The Baum-Welch algorithm is the most widely adopted methodology for model parameters estimation. The model parameters p_i, a_{ij}, mean μ_i and covariance $\sum_i$ are each characterized as:

$$\pi = \gamma_1(i) \quad (7)$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (8)$$

$$\mu_i = \frac{\sum_{t=1}^{T} \gamma_t(i) o_t}{\sum_{t=1}^{T} \gamma_t(i)} \quad (9)$$

$$\sum_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(o_t - \mu_i)(o_t - \mu_i)'}{\sum_{t=1}^{T} \gamma_t(i)} \quad (10)$$

where $\gamma_t(i)$ denotes the conditional probability of being in state *i* at time *t*, given the observations, and $\xi t(i, j)$ is the conditional probability of a transition from state i at time t to state j at time *t* + 1, given the observations. Both $\gamma t(i)$ and $\xi t(i, j)$ can be solved in terms of a well-known forward-backward algorithm [6]. Define the forward probability $\alpha_t(i)$ as the joint probability of observing the first *t* observation sequence O_1 *to* $t=\{o_1, o_2, \ldots, o_t\}$ and being in state *i* at time *t*. The $\alpha_t(i)$ can be solved inductively by the following formula:

$$\alpha_1(i) = \pi_i b_i(o_1), \ 1 \le i \le N \quad (11)$$

$$\alpha_{t+1}(i) = b(o_{t+1}) \sum_{j=1}^{N} [\alpha_t(i)\, \alpha_{ij}], \text{ For } 1 \le t \le T, \text{ For } 1 \le i \le N \quad (12)$$

Let the backward probability $b_t(i)$ be the conditional probability of observing the observation sequence $O_{t \text{ to } T}=\{o_{t+1}, o_{t+2}, \ldots, o_T\}$ after time t given that the state at time t is i. As with the forward probability, the $b_t(i)$ can be solved inductively as:

$$\beta_T = 1, \ 1 \le i \le N \quad (13)$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}\, b_j(o_{t+1}) \beta_{t+1}(j), \ t=T-1, T-2, \ldots, 1, \ 1 \le i \le N \quad (14)$$

The probabilities $\gamma t(i)$ and $\xi t(i, j)$ are then solved by:

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \quad (15)$$

$$\xi_t(i,j) = \frac{\alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i)\, a_{ij}\, b_j(o_{t+1})\beta_{t+1}(j)} \quad (16)$$

As a result, if both the observation density $b_i(o_t)$ and observation sequence $O=\{o_1, o_2, \ldots, o_T\}$ are well managed, then the hidden state sequence will be closer to the ideal situation. Moreover, the Viterbi algorithm is usually employed to perform global decoding which found the states of each observation separately. Using the Viterbi algorithm is aimed to find the most likely sequence of latent states corresponding to the observed sequence of data [7]. Also the Viterbi algorithm can be used to find the single best state sequence of an observation sequence. The Viterbi algorithm is another trellis algorithm which is very similar to the forward algorithm, except that the transition probabilities are maximized at each step, instead of summed[8].First define:

$$\delta_t(i) = \max_{q_1, q_2, \cdots, q_{t-1}} P(q_1\, q_2 \cdots q_t = s_i,\, o_1,\, o_2 \cdots o_t \,|\, \lambda) \quad (17)$$

as the probability of the most probable state path of the partial observation sequence. The Viterbi algorithm steps can be stated as:

1. Initialization

$$\delta_1(i) = \pi_i\, b_i(o_1), \ 1 \le i \le N, \quad (18)$$

2. Recursion:

$$\delta_t(j) = \max_{i=1}^{N} [\delta_{t-1}(i)\, a_{ij}]\, b_j(o_t), \ 2 \le t \le T, 1 \le j \le N \quad (19)$$

$$\Psi_t(j) = arg \max_{i=1}^{N} [\delta_{t-1}(i)\, a_{ij}], \ 2 \le t \le T, \ 1 \le j \le N \quad (20)$$

3. Termination:

$$P^* = \max_{i=1}^{N} [\delta_T(i)] \quad (21)$$

$$q^*_T = arg \max_{i=1}^{N} [\delta_T(i)] \quad (22)$$

When implementing HMM for unsupervised image classification, the pixel values (or vectors) correspond to the observations, and after the estimation of the model parameter is completed, the hidden state then corresponds to the cluster to which the pixel belongs.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

III. CLASSIFICATION

The main purpose of satellite and other imagery classification is the recognition of objects on the Earth's surface and their presentation in the form of thematic maps. Land cover is determined by the observation of grey values in the imagery. Classification is one of the most important steps in handling remote sensing imagery and represents important input data for geographic information systems [9]. There are two types of land-cover classification, supervised and unsupervised classification. Supervised image classification relies on statistical parameters of a class generated during training sampling which is in general nontransferable from one image to the other. Unsupervised classification with a clustering technique provides automated grouping, but there is no way to establish a fixed relation between a cluster code and a certain land cover category [10]. Cases of unsupervised classification, some of the statistical properties of the different classes are unknown and have to be estimated with iterative methods such as estimation-maximization (EM) [11]. ISODATA, K-means clustering algorithms are used for unsupervised classification [13]. While maximum likelihood, minimum distance, and mahalanobis distance are the most popular methods used in supervised classification [12].

IV. RELATED RESEARCHES

The literature discusses different approaches for the remotely sensed imagery classification. This survey will summarizes the relevant remotely sensed imagery classification algorithms in last years.

An application of the HMM in hyperspectral image analysis has been introduced by Du and Chang (2001). This application inspired by the analogy between the temporal variability of a speech signal and the spectral variability of a remote sensing image pixel vector, that models a hyperspectral vector as a stochastic process, where the spectral correlation and band-to-band variability are modeled by a hidden Markov process with parameters determined by the spectrum of the vector that forms a sequence of observations. With this interpretation, a new HMM based spectral measure, referred to as the HMM information divergence (HMMID), is derived to characterize spectral properties. The performance of this new measure was evaluated by comparing it with three commonly used spectral measures, Euclidean distance (ED) and the spectral angle mapper (SAM), and the recently proposed spectral information divergence (SID). The experimental results show that the HMMID performs better than the other three measures in characterizing spectral information at the expense of computational complexity [13].

Chen and Stow (2003) proposed other strategies for integrating information from multiple spatial resolutions into land-use/land-cover classification routines. They presents three strategies for selecting and integrating information from different spatial resolutions into classification routines. One strategy is to combine layers of images of varying resolution. A second strategy involves comparing the posteriori probabilities of each class at different resolutions. Another strategy was based on a top-down approach starting with the coarsest resolution. The classification accuracy obtained from using three multiple strategies was greater when compared with that from a conventional single-resolution approach. Among the three strategies, the top-down approach resulted in the highest classification accuracy with a Kappa value of 0.648, compared to a Kappa of 0.566 for the conventional Classifier [14].

A developed unsupervised classification approach was introduced by Tso and Olsen (2005). This approach was based on observation-sequence and observation-density adjustments, which have been proposed for incorporating 2D spatial information into the linear HMM. For the observation-sequence adjustment methods, five neighbourhood systems have been proposed. Two neighbourhood systems were incorporated into the observation density methods. The classification accuracy then evaluated by means of confusion matrices made by randomly chosen test samples. Experimental results showed that the proposed approaches for combining both the spectral and spatial information into HMM unsupervised classification mechanism present improvements in both classification accuracy and visual qualities [4].

Another group proposed a method (2007) to model temporal knowledge and to combine it with spectral and spatial knowledge within an integrated fuzzy automatic image classification framework for land-use land-cover map update applications. The classification model explores not only the object features, but also information about its class at a previous date. The method expresses temporal class dependencies by means of a transition diagram, assigning a possibility value to each class transition. A Genetic Algorithm (GA) carries out the class transition possibilities estimation. Temporal and spectral/spatial classification results were combined by means of fuzzy aggregation. The experiments showed that the use of temporal knowledge markedly improved the classification performance, in comparison to a conventional single-time classification. A further observation was that multitemporal knowledge might subsume the knowledge related to steady spatial attributes whose values do not significantly change over time [15].

Li and Zhang (2011) introduces an expert interpretation-based Markov chain geostatistical (MCG) framework for classifying Land-Use/Land-Cover (LULC) classes from remotely sensed imagery. The framework uses the MCG method to classify uninformed pixels based on the informed pixels and quantify the associated uncertainty. The method consists of four steps:

1) Decide the number of LULC classes and define the physical meaning of each class.
2) Obtain a data set of class labels from one or a time series of remotely sensed images through expert interpretation.
3) Estimate transiogram models from the data set.
4) Use the Markov chain sequential simulation algorithm to conduct simulations that are conditional to the data set.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

The simulated results not only provide classified LULC maps but also quantify the uncertainty associated with the classification. Although it is relatively labor intensive, such an expert interpretation and geostatistical
Simulation-based approach may provide a useful LULC classification method complementary to existing image processing methods, which usually account for limited expert knowledge and may not incorporate ground observation data or assess the uncertainty associated with classified data [15].

A method for land cover land use classification with TWOPAC (TWinned Object and Pixel based Automated classification Chain) was developed by Huth, Kuenzer, Wehremann, Gebhardt, Tuan, and Deeh (2012) this method enables the standardized, independent, user-friendly, and comparable derivation of LC and LU information, with minimized manual classification labor. TWOPAC allows classification of multi-spectral and multi-temporal remote sensing imagery from different sensor types. TWOPAC enables not only pixel-based classification, but also allows classification based on object characteristics. Classification is based on a Decision Tree approach (DT) for which the well-known C5.0 code has been implemented, which builds decision trees based on the concept of information entropy. TWOPAC enables automatic generation of the decision tree classifier based on a C5.0-retrieved ascii-file, as well as fully automatic validation of the classification output via sample based accuracy assessment. Envisaging the automated generation of standardized land cover products, as well as area-wide classification of large amounts of data in preferably a short processing time, standardized interfaces for process control, Web Processing Services (WPS), as introduced by the Open Geospatial Consortium (OGC), are utilized. TWOPAC's functionality to process geospatial raster or vector data via web resources (server, network) enables TWOPAC's usability independent of any commercial client or desktop software and allows for large-scale data processing on servers. Furthermore, the components of TWOPAC were built-up using open source code components and were implemented as a plug-in for Quantum GIS software for easy handling of the classification process from the user's perspective [17].

Beulah and Tamilarasi (2012) developed a technique which tries to extract useful information from large set of satellite images. They consider time series satellite images for their research. Image processing techniques used enhance the satellite images. They tried to analyze land covers of a particular area and extract information about their vegetation. Time series satellite images were analyzed with scalable and efficient methods. Last step the required features were extracted based on their texture and the vegetation features were collected [18].

V. THE PROPOSED ALGORITHM

This paper proposes an unsupervised Hidden Markov Models (HMM) to classify the land cover from multispectral satellite images and propose a new technique of sequencing the pixels of image in order to form the observation vector in a way that can fit into the linear HMM. The schemes of incorporating 2D spatial information of remotely sensed imagery into a one-dimensional linear HMM have been proposed and demonstrated in terms of accuracy analysis and visual quality through unsupervised classifications.

REFERENCES

[1] Mohamed El Yazid Boudaren, and Abdel Belaid, "A New scheme for land cover classification in aerial images: combining extended dependency tree-HMM and unsupervised segmentation," Lecture Notes in Electronic Engineering - Electronic Engineering and Computing Technology Spriger, inroad-00579704, version 1-24, pp. 471-482, Mar 2011.

[2] Qian Wang, Jianping Chen, and Yi Tian, "Remote sensing image interpretation study serving urban planning based on GIS," The international Archives of the Photogrammetry, Remote sensing and Spatial Information Sciences. Vol. XXXVII. Part B4. Beijing 2008.

[3] Koray Kayabol, and Josiane Zerubia, "Unsupervised mplitude and texture based classification of SAR images with multinomial latent model," Institute National de Recharge en Information en Automatique. Hal-00612491, version 2-2 May 2012.

[4] Tso B. and Olsen R. C., "Combining spectral and information into hidden markov models for unsupervised image classification," International Journal of Remote Sensing. Vol. 26, No. 10, pp. 2113-2133, 20 May 2005.

[5] Wleed Abdulla, and Nikola Kasabov, "The concept of hidden markov model in speech recognition," The Information Science Discussion Paper Series, No. 99/09, ISSN 1177-45X, May 1999.

[6] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss, "A Maximization technique occurring in the statistical analysis of probabilistic functions of markove chains," The Annals of Mathematical statistics, Vol. 41, No. 1, pp. 164-171, 1970.

[7] Silvia Pandolfi, and Francesco Bartolucci, "A new constant memory recursion for hidden markov models," FIRB ("Future in ricerca" 2012), Perugia (IT), March 15-16, 2013.

[8] Daniel Jurafsky, and Jamse H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics and speech recognition," 2nd Ed., prentice-Hall 2000, ISBN: 0-13-095069-6.

[9] Ziga Kokalj, Kristof Ostir, "Land covermapping using landsat satellite image classification in the classical karst – kras region," ACTA Carsologica 36/3, pp. 433-440, Postojna, 2007.

[10] Nguyen Dinh Duong, "Land cover category definition by image invariants for automated classification," International archives of photogrammetry and remote sensing, Vol. XXXIII, Part B7, pp. 985-991, Amsterdam 2000.

[11] Roger Fjørtoft, Jean-Marc Boucher, Yves Delignon, Rene Garello, Jean-Marc Le Caillec, Henri Maître, Jean-Marie Nicolas, Wojciech Pieczynski, Marc Sigelle, and Florence Tupin,"Unsupervised classification of radar images based on hidden markov models and generalised mixture estimation," Proc. SAR Image Analysis Modeling, and Techniques V, Vol.SPIE 4173, Barcelona, Spain, September 2000.

[12] F.S. Al-Ahmadi, A.S. Hames, "Comparision of four classification methods to extract land use and land cover from raw satellite images for some remote arid areas, Kingdom of

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Saudi Arabia," JKAU, Earth Sci., Vol.20, No. 1, pp. 167-191, 2009.

[13] Qian Du, Chein-l Chang, "Hidden Markov Model approach to spectral analysis for hyperspectral imagery," Society of Photo-Optical Instrumentation Engineers, Opt. Eng. 40(10), pp. 2277-2284, October 2001.

[14] DongMei Chen, Douglas Stow, "Strategies for integrating information from multiple spatial resolutions into land-use/land-cover classification routines," Photogrammetric Engineering & Remote sensing, Vol. 69, No. 11, pp. 1279-1287, November 2008.

[15] Guilherme L.A. Mota, Raul Q. Feitosa, Heitor L.C. Coutinho, Claus-Eberhard Liedtke, Sönke Müller, Kian Pakzad, Margareth S.P. Meirelles, "Multitemporal fuzzy classification model based on class transition possibilities," ISPRS Journal of Photogrammetry & Remote sensing 62, pp. 186-200, 2007.

[16] Weidong Li and Chuanrong Zhang, "A markov chain geostatistical framework for land-cover classification with uncertainty assessment based on expert-interpreted pixel form remotely sensed imagery," IEEE transactions on geoscience and remote sensing, Vol. 49, No. 8, 2983-2992, August 2011.

[17] Juliane Huth, Claudia Kuenzer, Thilo Wehrmann, steffen Gebhardt, Vo Quoc Tuan, and Stefan Dech, "Land cover and land use classification with TWOPAC: towards automated processing for pixel- and object-based image classification," Remote Sens. 4, doi:10.3390/rs4092530, pp. 2530-2553, 2012.

[18] Beulah J., and Tamilarasi M., "Processing and analysis of satellite images to detect vegetation," International Journal of Communications and Engineering, Vol. 03, No. 3, Issue:03, pp. 70-73, March 2012.

Secure Routing in UAV

Ahmed Refaat Sobhy
Arab Academy for Science
& Technology & Maritime Transport
College of Engineering & Technology
Cairo, Egypt

Rowayda.A.Sadek
Arab Academy for Science
& Technology & Maritime Transport
College of Engineering & Technology
Cairo, Egypt

Atalla Hashad
Arab Academy for Science
& Technology & Maritime Transport
College of Engineering & Technology
Cairo, Egypt

***Abstract*— The field of UAV has gained an important part of the interest of researchers and become very popular in last few years. Focusing in the routing protocols used in UAV's systems in order to obtain a secure routing protocol this paper presents the effect of DOS attack on two different types of routing protocols , proactive and reactive routing protocols. The proactive routing protocol is represented by OLSR routing protocol and the reactive routing protocol is represented by AODV , TORA routing protocols . in addition the performance metrics of ordinary routing protocols (OLSR , AODV , TORA) are compared in three different scenarios implemented by using Opnet simulator. The simulation results will show the performance impact of security implements into reactive & proactive protocols after implementations of Authentication & encryption algorithms. The main concern of this paper is to propose an efficient and secure routing protocol to the UAV.**

I. INTRODUCTION

Unmanned Aerial Vehicle (UAV), which is an aircraft with no pilot on board. **UAVs** can be remote controlled aircraft or can fly autonomously based on pre-programmed flight plans or more complex dynamic automation systems. UAVs are currently used for a number of missions, including reconnaissance and attack roles.

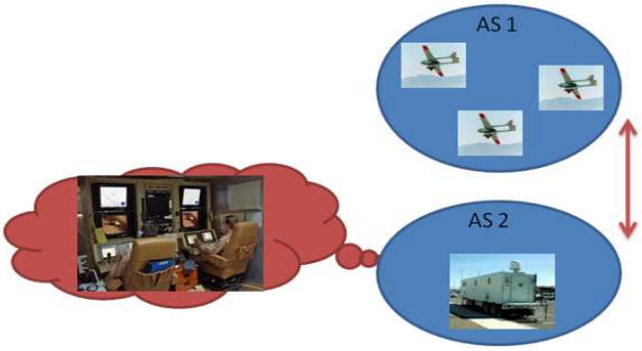


Fig.1 Unmanned Aerial Vehicle System

Fig 1 shows UAV system which is described as a two autonomous systems, first autonomous system; AS1includes the UAVs and second autonomous system; AS2 includes the base station which includes racks where the operators tracks the UAVs.

When using UAV, the communication network is an important aspect to be considered. Assuming that the communication of these devices will be wireless, there are many problems involving this technology. The communication of these UAVs with the operational base station is a very important issue because of the critical information that has to be transmitted. In other hand the routing between UAVs and their base station must be secured. The protocols that run complex military systems such as UAVs can be easily hacked or destroyed by malicious nodes but it is required for military systems to be run by fail-safe protocols. One aspect of UAV networks that complicates the design of a secure routing protocol is in-network aggregation [17]. In networks that are more conventional a secure routing protocol is typically only required to guarantee message availability[18].

Message integrity and confidentiality are handled at a higher layer by an end-to-end security mechanism such as SSH or SSL [2]. Ideal system should guarantee the confidentiality, integrity, authenticity and availability of all messages in the presence of resourceful adversaries. In such an environment, there is no guarantee that a path between two nodes would be free of malicious nodes that would not comply with the employed protocol and would attempt to harm the network operation. Secure routing protocols have emerged in recent years [2]. However, most of them assume an insecure environment in which no node can be trusted, and adversary nodes can be inside the network. Researchers try to put every insecure factor into consideration and propose a perfectly secure scheme. Absolute security is impossible to obtain[19]. Thus, studying the performance of secure routing protocols in malicious environments is needed in order to secure routing protocols in malicious environments. In this paper three scenarios are implemented by using OPNET simulator, the first scenario shows three different routing protocols (OLSR, AODV, TORA) without hacking (Normal) & the second scenario by implementing hacking using DOS attack by explicitly generating data packets based on real-life TCP Dump data that contain intrusion packets [1]. The third scenario is carried out by securing the routing protocols using Authentication & encryption algorithm. The rest of this paper is organized as follows: Section 2 building scenario for AODV, OLSR, TORA without hacking (Normal scenario), Section 3 hacking the scenarios using DOS attack, Section 4 securing routing protocols (AODV, OLSR, TORA) using encryption & authentication algorithms, Section 5 simulation analysis and finally section 6 concludes the paper.

II. NORMAL SCENARIOS WITHOUT HACKING

The proposed UAV based system scenario is represented by using OPNET simulator as shown in fig.2. UAVs are represented by six nodes (U1 to U6) which represents

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

autonomous system called 200, the main base station is represented by a node connected to Rack1 and Rack 2 representing autonomous system called 100. The altitude and the distance are neglected .

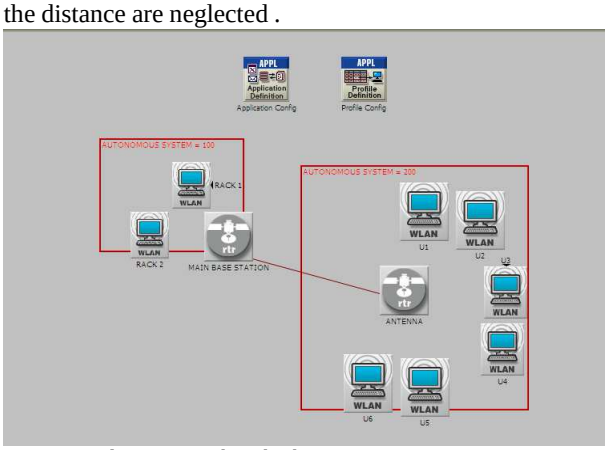


Fig.2 Normal scenario without hacking

The scenario implements a mission for six UAVs ,the UAVs sends a real video stream to the main base station to be seen in Rack1 and Rack2, also the base station sends data to the UAVs in order to control the UAVs and guidance them . The data measured are wireless lan delay, wireless lan load, routing traffic sent & received (pkts/sec), Throughput(bits/sec), Traffic dropped (pkts/sec)[20].

A. AODV scenario without hacking

The AODV routing protocol [3-4], is a pure on-demand routing protocol. One of the distinguished features of AODV is its use of a destination sequence number of each route entry [5].The AODV algorithm enables dynamic, self-starting, multi-hop routing between participating mobile nodes wishing to establish and maintain an ad hoc network. Several attacks, can be launched against AODV routing protocol such as message tampering attack, message dropping attack and message replay attack[6,7]. As shown in Fig.3, the routing traffic in the whole network is high and this is due to the high connectivity between UAVs & the base station also there is no data dropped or lost and this can be seen in fig.8 & numerically the base station average is 10.966 pkts/sec , rack1 average is 5.262 pkts/sec , rack2 average is 5.061 pkts/sec. From fig.4 due to the high command sent by the base station to the UAVs we found that the amount of data received by the UAVs is high and numerically UAV 5 average is 23.711 pkts/sec, rack 2 average is 12.622 pkts/sec, rack1 average is 10.861 pkts/sec.

From fig.5 due to the high amount of data and commands sent and received between UAVs we found that the wireless lan delay (sec) for UAVs are high & numerically UAV 6 average is 0.34847 sec, rack 2 average is 0.00762 sec , rack1 average is 0.00941 sec. From fig.6 due to wireless lan delay (sec), high amount of data and commands sent and received between UAVs we found that the wireless lan load for UAVs is high & numerically UAV 5 average is 221.347.433 bits/sec, the base station average is 1.589.826 bits/sec. From fig .7 the wireless lan through put is the average rate of successful message delivery over a communication channel and we found that numerically the base station average is 10,421,479 bits/sec & UAV2 average is 53.616 bits/sec. From fig.8 no packet dropped when applying AODV in normal scenario.

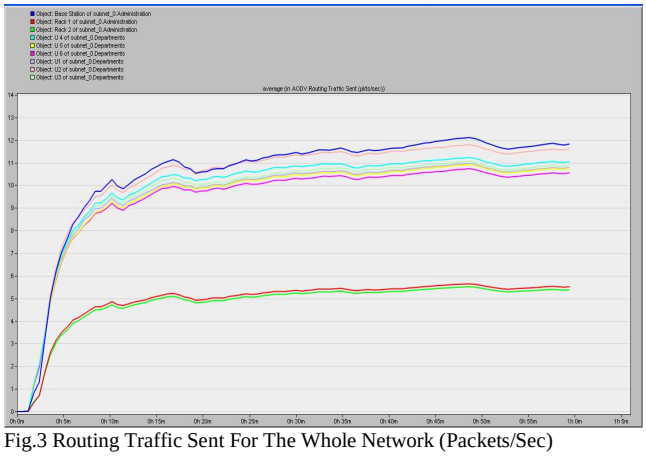


Fig.3 Routing Traffic Sent For The Whole Network (Packets/Sec)

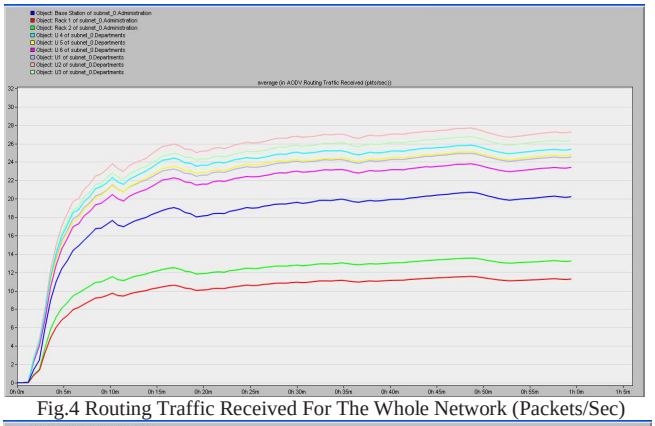


Fig.4 Routing Traffic Received For The Whole Network (Packets/Sec)

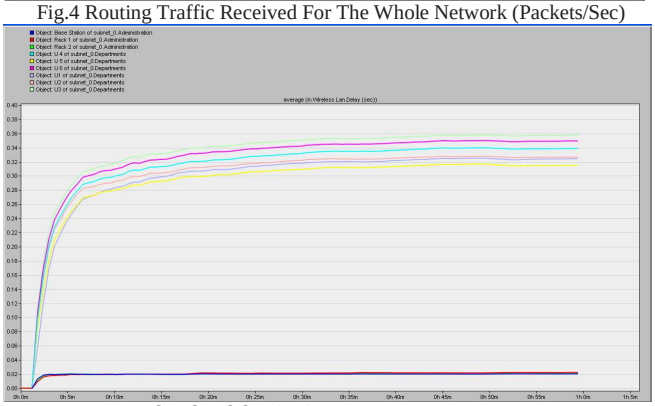


Fig .5 wireless lan delay (sec)

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

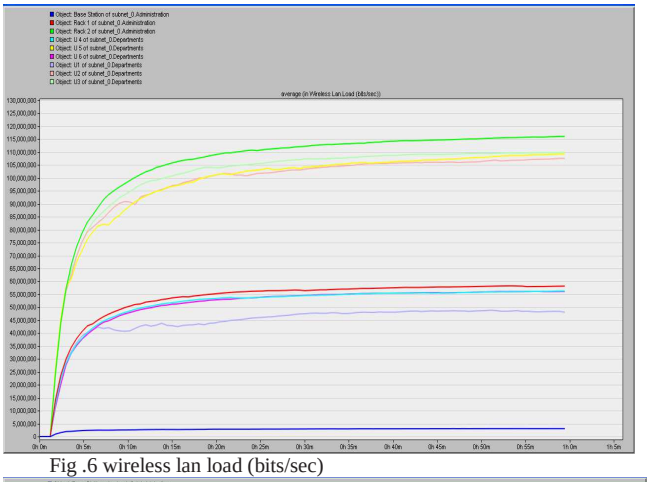


Fig .6 wireless lan load (bits/sec)

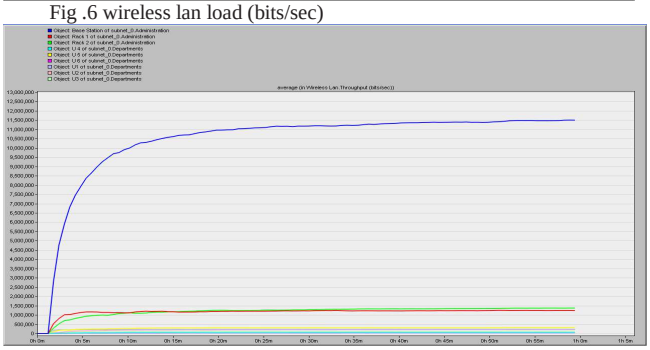


Fig .7 wireless lan through put (bits/sec)

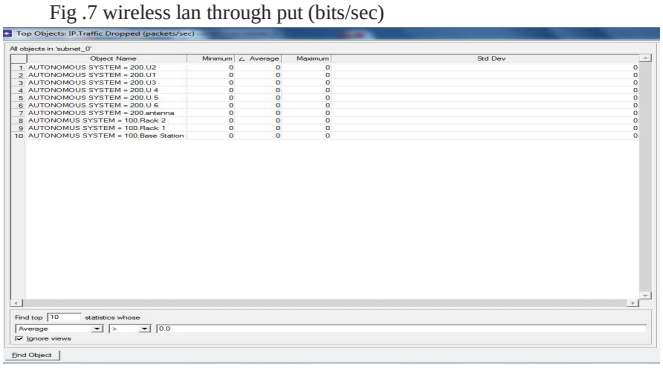


Fig .8 traffic dropped (pkts/sec)

B. OLSR scenario without hacking

OLSR routing protocol inherits the stability of the link state algorithm. Due to its proactive nature [8], it has an advantage of having the routes immediately available when needed. In a pure link state protocol, all the links with neighbor nodes are declared and are flooded in the entire network. OLSR protocol is an optimization of a pure link state protocol for mobile ad hoc networks. First, it reduces the size of control packets: instead of all links, it declures only a subset of links with its neighbors who are its multipoint relay sectors.

Secondly, it minimizes flooding of this control traffic by using only the selected nodes, called multipoint relays, to diffuse its messages in the network. Only the multipoint relays of a node retransmit its broadcast message. This technique significantly reduces the number of retransmissions in a flooding or broadcast procedure[9,10].

A part from normal periodic control messages, the protocol does not generate extra control traffic in response to link failures and additions. The protocol keeps the routes for all the destinations in the network. The protocol is designed to work in a completely distributed manner and thus does not depend upon any central entity. The protocol does not require a reliable transmission for its control messages: each node sends its control messages periodically, and can therefore sustain a loss of some packets from time to time, which happens very often in radio networks due to collisions or other transmission problems. The protocol also does not need an in-order delivery of its messages: each control message contains a sequence number of most recent information; therefore the re-ordering at the receiving end cannot make the old information interpreted. OLSR uses two kinds of the control messages: Hello and Topology Control (TC). Hello messages are used for finding the information about the link status and the host's neighbors [11]. With the Hello message the Multipoint Relay (MPR) Selector set is constructed which describes which neighbors has chosen this host to act as MPR and from this information the host can calculate its own set of the MPRs. Implementing the same scenario by using OLSR protocol without hacking as shown in fig.9 and obtaining the same data measured for AODV to be compared.

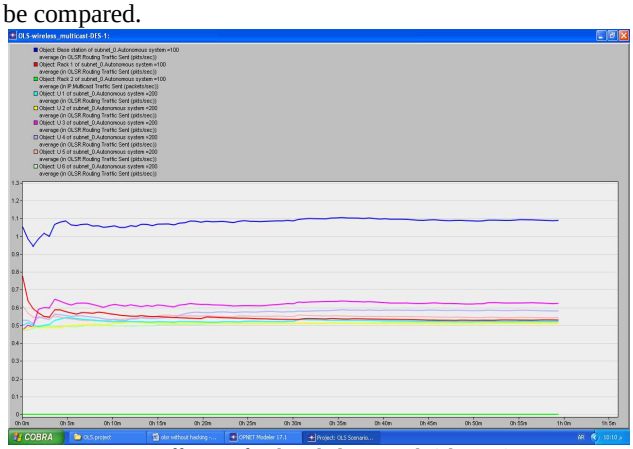


Fig .9 routing traffic sent for the whole network (pkts/sec)

From fig.9 the higher is the base station & the lower is UAV 6,2 in routing traffic sent & numerically the base station average is 1.0894 pkts/sec , UAV 6 average is 0.5178 pkts/sec , UAV 2 average is 0.5119 pkts/sec.

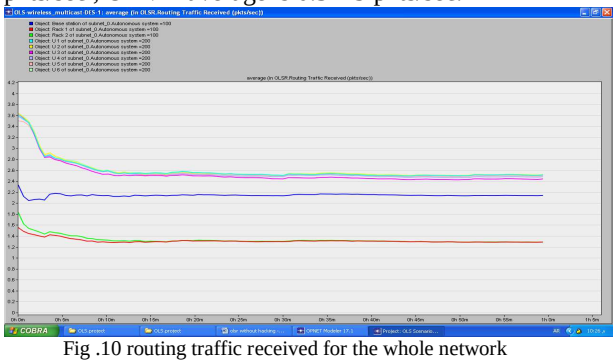


Fig .10 routing traffic received for the whole network

From fig.10 the higher is UAV 2 & the lower is rack 1 in routing traffic received & numerically UAV 2 average 2.5194

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

pkts/sec , rack 2 average is 1.2892 pkts/sec , rack1 average is 1.2886 pkts/sec.

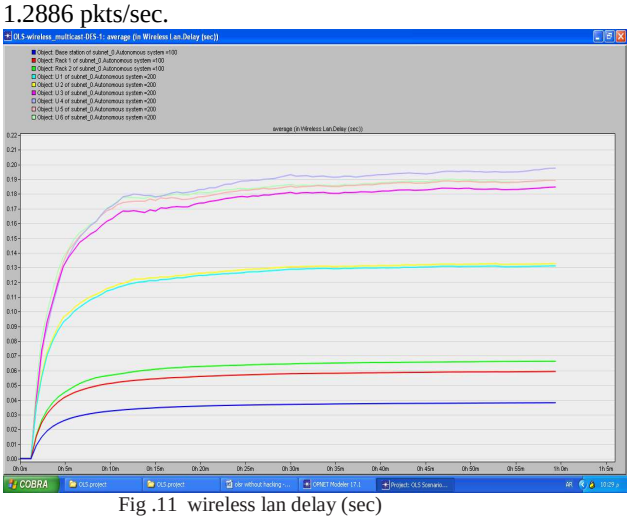


Fig .11 wireless lan delay (sec)

From fig.11 the higher is UAV 4 & the lower is base station in wireless lan delay & numerically UAV 4 average is 0.19757 sec, rack 1 average is 0.05937 sec , base station average is 0.03809 sec.

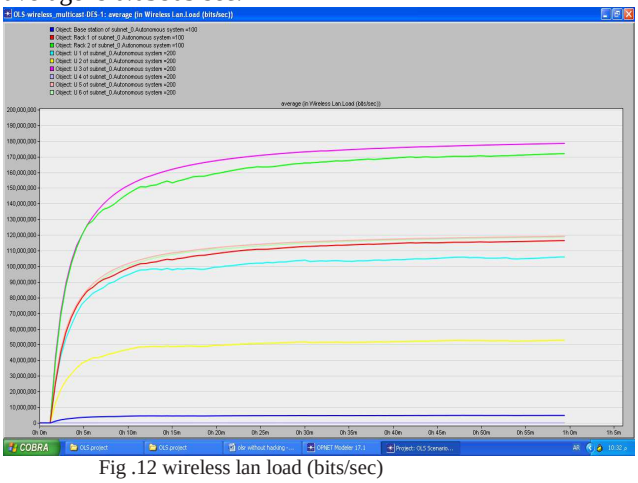


Fig .12 wireless lan load (bits/sec)

From fig.12 the higher is UAV 3 & the lower is UAV 4 in wireless lan load & numerically UAV 3 average is 178.482.204 bits/sec, UAV 4 average is 15.344 bits/sec.

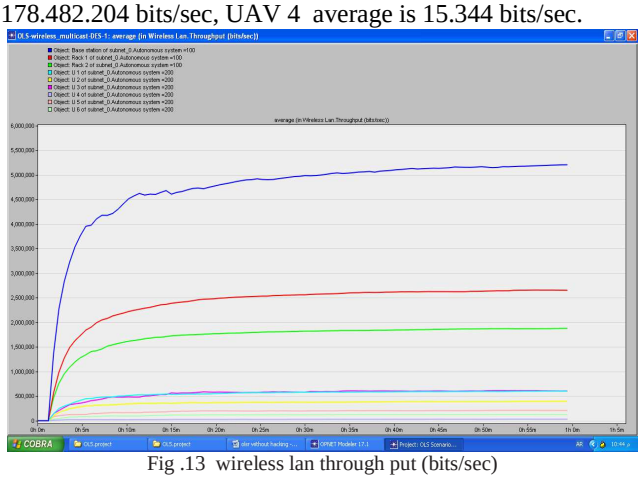


Fig .13 wireless lan through put (bits/sec)

From fig .13 the higher is the base station & the lower is UAV4 and numerically the base station average is 5,206,587 bits/sec & UAV4 average is 32.972 bits/sec.

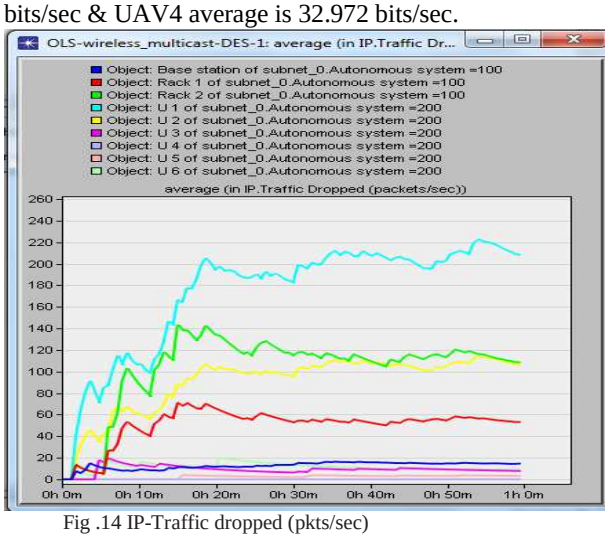


Fig .14 IP-Traffic dropped (pkts/sec)

From fig 14 the higher is UAV1 with 208.33 average pkts/sec and the lower is UAV4 with 0.03 average pkt/sec.

C. TORA scenario without hacking

TORA is adaptive and scalable routing algorithm based on the concept of link reversal. It finds multiple routes from source to destination in a highly dynamic mobile networking environment[11].An important design concept of TORA is that control messages are localized to a small set of nodes nearby a topological change. Nodes maintain routing information about their immediate one-hop neighbors. The protocol has three basic functions: route creation, route maintenance, and route erasure . Nodes use a “height” metric to establish a directed cyclic graph (DAG) rooted at the destination during the route creation and route maintenance phases. The link can be either an upstream or downstream based on the relative height metric of the adjacent nodes. TORA’s metric contains five elements: the unique node ID, logical time of a link failure, the unique ID of a node that defined the new reference level, a reflection indicator bit, and a propagation ordering parameter. Establishment of DAG resembles the query/reply process discussed in Lightweight Mobile Routing (LMR). Route maintenance is necessary when any of the links in DAG is broken. The main strength of the protocol is the way it handles the link failures. TORA’s reaction to link failures is optimistic that it will reverse the links to re-position the DAG for searching an alternate path. Effectively, each link reversal sequence searches for alternative routes to the destination. This search mechanism generally requires a single-pass of the distributed algorithm since the routing tables are modified simultaneously during the outward phase of the search mechanism. Other routing algorithms such as LMR use two-pass whereas both DSR and AODV use three pass procedure. TORA achieves its single-pass procedure with the assumption that all the nodes have synchronized clocks (via GPS) to create a temporal order of topological change of events. The “height” metric is dependent on the logical

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

time of a link failure [12,13,14]. The advantage of TORA is that the multiple routes are supported by this protocol between the source and destination node. Therefore, failure or removal of any of the nodes is quickly resolved without source intervention by switching to an alternate route to improve congestion. It does not require a periodic update, consequently communication overhead and bandwidth utilization is minimized. It provides the support of link status sensing and neighbor delivery, reliable in-order control packet delivery and security authentication. Implementing the same scenario by using TORA protocol without hacking and obtaining the same data measured for AODV & OLSR to be compared.

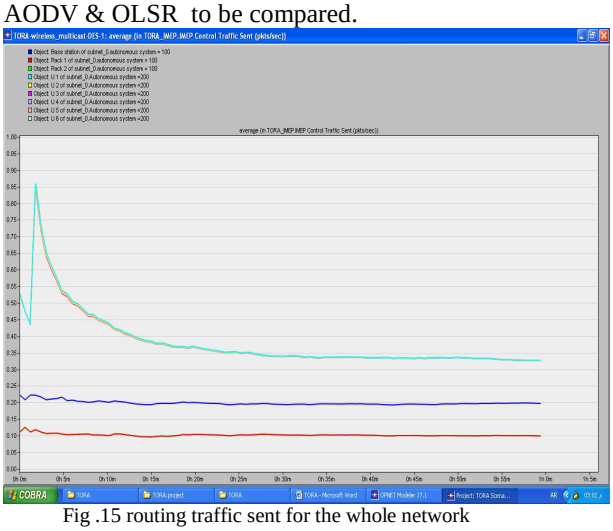


Fig .15 routing traffic sent for the whole network

From fig.15 the higher is UAV1 & the lower is Rack 2,1 in routing traffic sent & numerically UAV1 average is 0.32694 pkts/sec , Rack 2 average is 0.09917 pkts/sec , Rack1 average is 0.09917 pkts/sec.

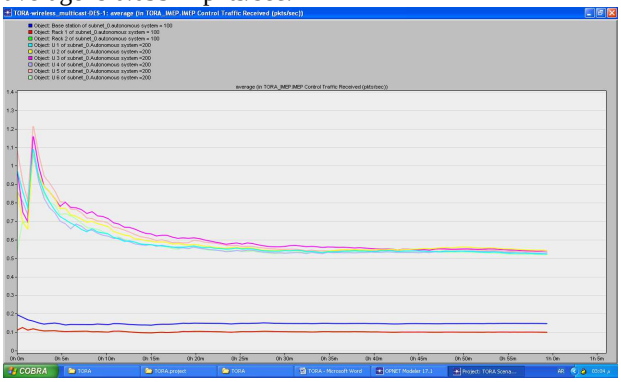


Fig .16 routing traffic received for the whole network

From fig.16 the higher is UAV 2 & the lower is rack2, 1 in routing traffic received & numerically UAV 2 average is 0.54111 pkts/sec , rack 2 average is 0.9917 pkts/sec , rack1 average is 0.9917 pkts/sec.

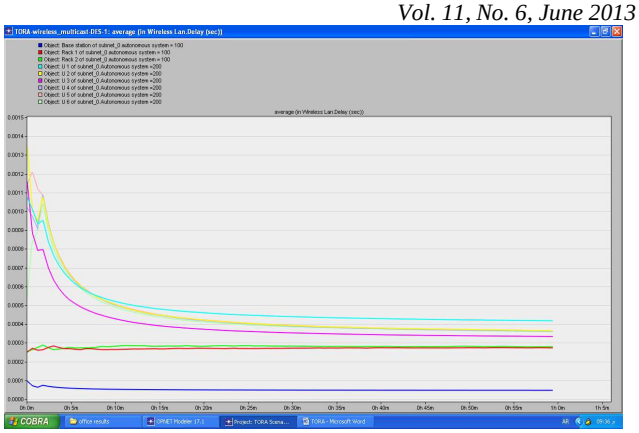


Fig .17 wireless lan delay (sec)

From fig.17 the higher is UAV 1 & the lower is base station in wireless lan delay & numerically UAV 1 average is 0.00041758 sec, base station average is 0.00004741 sec.

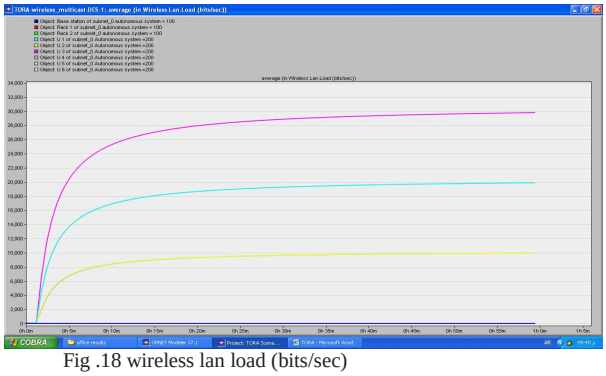


Fig .18 wireless lan load (bits/sec)

From fig.18 the higher is UAV 3 & the lower is rack 2 in wireless lan load & numerically UAV 3 average is 29.801 bits/sec, rack2 average is 16 bits/sec.

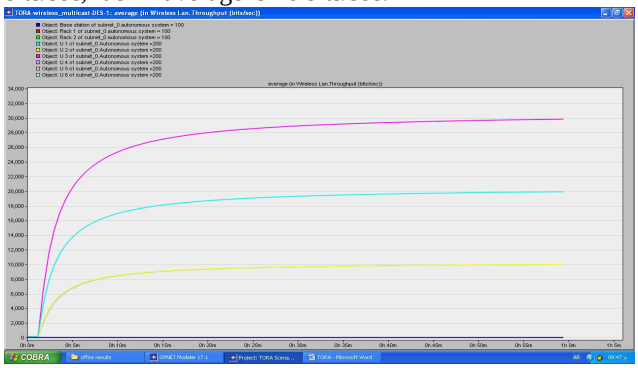


Fig .19 wireless lan through put (bits/sec)

From fig .19 the higher is UAV3& the lower is the base station and numerically the UAV3 average is 29.838 bits/sec &the base station average is 16 bits/sec.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

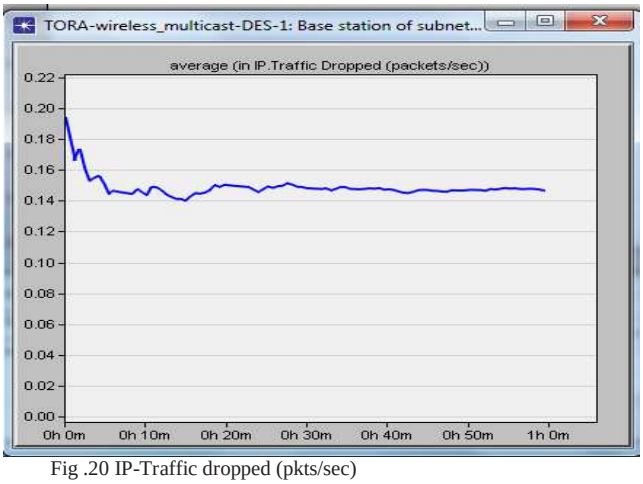


Fig .20 IP-Traffic dropped (pkts/sec)

From fig.20 and by ignoring the antenna results in the whole scenarios we find that the traffic dropped is only for the base station and its average is 0.14667 pkts/sec.

III. HACKING THE SCENARIOS USING DOS ATTACK

The DOS attack is the way of preventing legitimate users of a service or network resource from accessing that service or resource ,DOS attacks usually uses software bugs to freeze or crash the network resource or bandwidth limits by making use of a flood attack to saturate all bandwidth. Applying DOS attack for the last three scenarios by using file containing the process table attack packets, DARPA Intrusion Detection Evaluation project [1], and by getting the same results to be compared we get the following results.

A. AODV scenario under attack using DOS attack

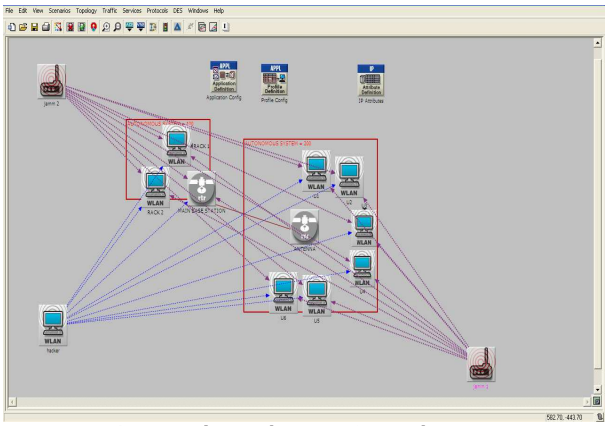


Fig.21 AODV under attack using DOS attack

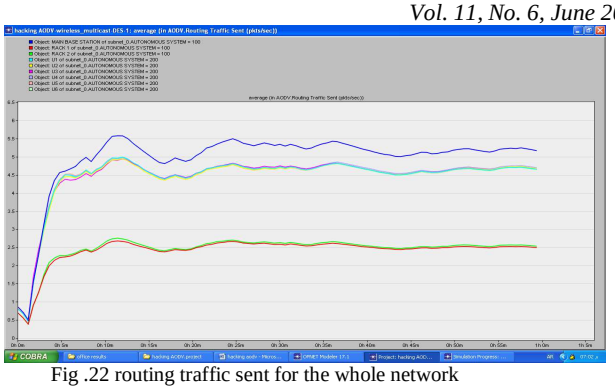


Fig .22 routing traffic sent for the whole network

From fig.22 the higher is the base station & the lower is Rack 2,1 in routing traffic sent and numerically the base station average is 7.9986 pkts/sec , Rack 2 average is 2.5333 pkts/sec , Rack1 average is 2.4942 pkts/sec.

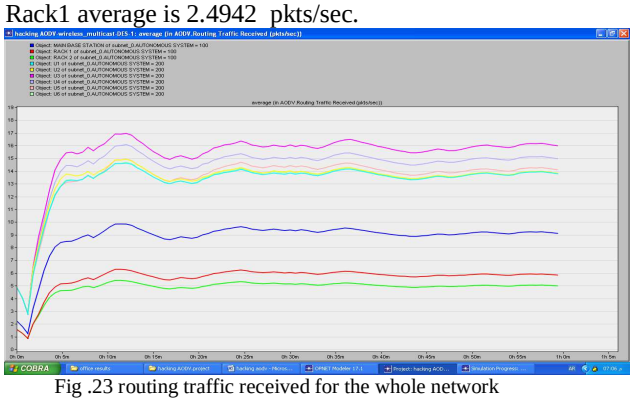


Fig .23 routing traffic received for the whole network

From fig.23 the higher is UAV 3 & the lower is rack1, 2 in routing traffic received & numerically UAV 3 average is 15.981 pkts/sec , rack 1 average is 5.835 pkts/sec , rack2 average is 4.985 pkts/sec.

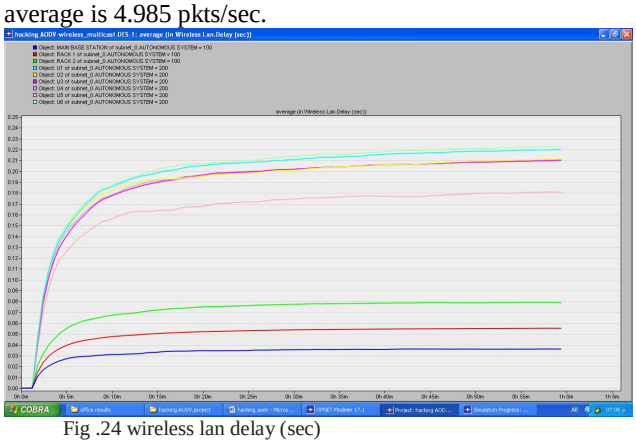


Fig .24 wireless lan delay (sec)

From fig.24 the higher is UAV 6 & the lower is base station in wireless lan delay & numerically UAV 6 average is 0.22293 sec, base station average is 0.03615 sec.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

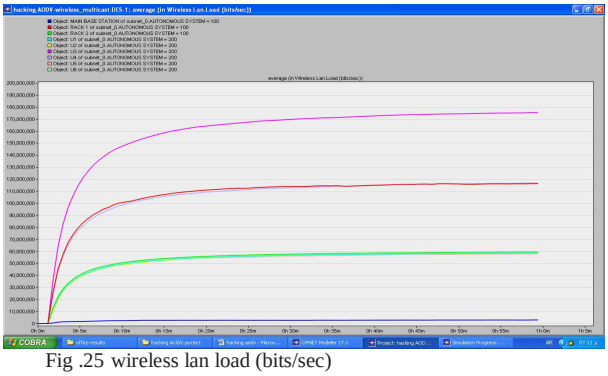


Fig .25 wireless lan load (bits/sec)

From fig.25 the higher is UAV 3 & the lower is the base station in wireless lan load & numerically UAV 3 average is 175.361.335 bits/sec, the base station average is 3.017.378 bits/sec.

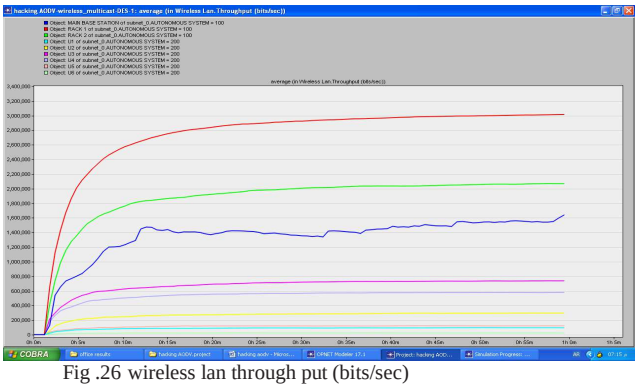


Fig .26 wireless lan through put (bits/sec)

From fig .26 the higher is Rack1& the lower is UAV6 and numerically the Rack1 average is 3,016,726 bits/sec & UAV6 average is 25.514 bits/sec.

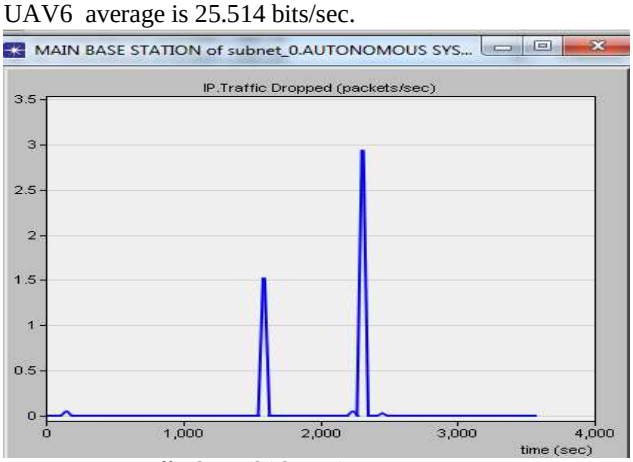


Fig .27 IP-Traffic dropped (pkts/sec)

From fig.27 the traffic dropped is only for the base station and its average is 0.046111 pkts/sec.

B. OLSR scenario under attack using DOS attack

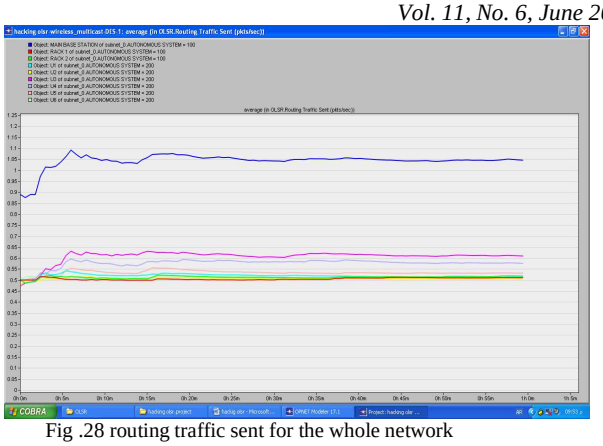


Fig .28 routing traffic sent for the whole network

From fig.28 the higher is the base station & the lower is UAV2 in routing traffic sent & numerically the base station average is 1.0450 pkts/sec , UAV2 average is 0.5075 pkts/sec.

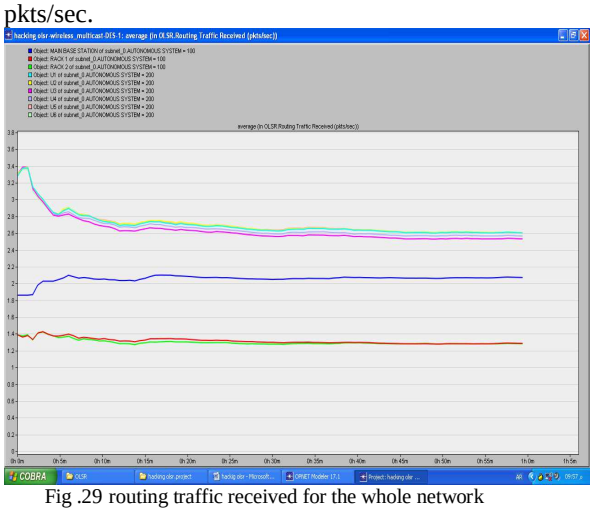


Fig .29 routing traffic received for the whole network

From fig.29 the higher is UAV 6 & the lower is rack1, 2 in routing traffic received & numerically UAV 6 average is 4.2908 pkts/sec , rack 1 average is 1.2864 pkts/sec , rack2 average is 1.2825 pkts/sec.

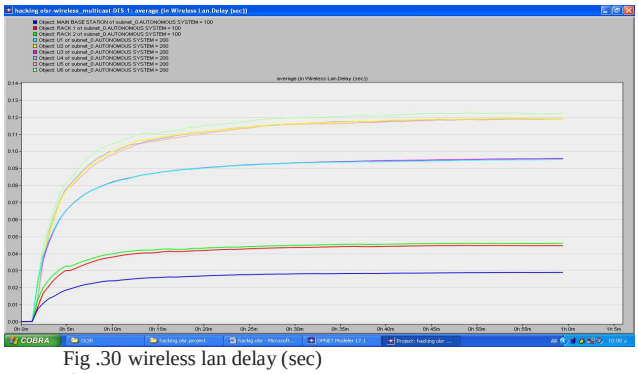


Fig .30 wireless lan delay (sec)

From fig.30 the higher is UAV 6 & the lower is base station in wireless lan delay & numerically UAV 6 average is 0.12224 sec, base station average is 0.02894 sec.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

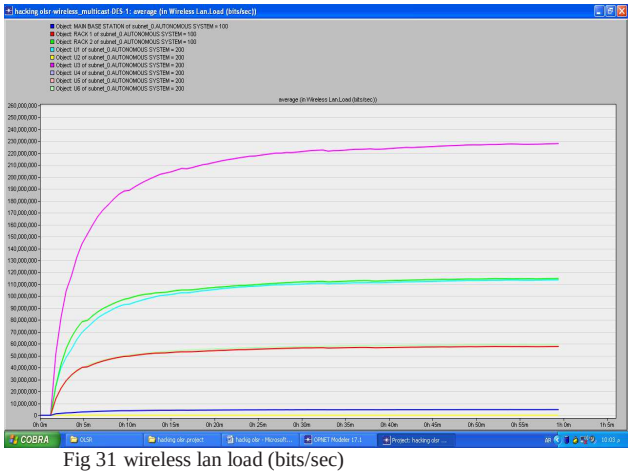


Fig 31 wireless lan load (bits/sec)

From fig.31 the higher is UAV 3 & the lower is the UAV2 in wireless lan load & numerically UAV 3 average is 228.149.644 bits/sec, UAV2 average is 9.513 bits/sec.

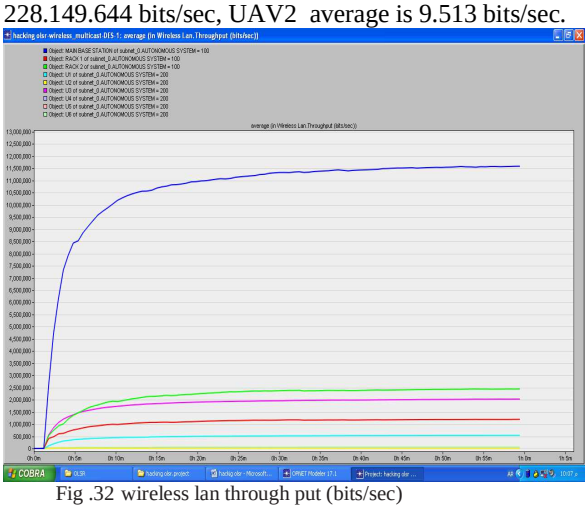


Fig .32 wireless lan through put (bits/sec)

From fig .32 the higher is the base station& the lower is UAV2 and numerically the base station average is 11,589,762 bits/sec & UAV2 average is 8.831 bits/sec.

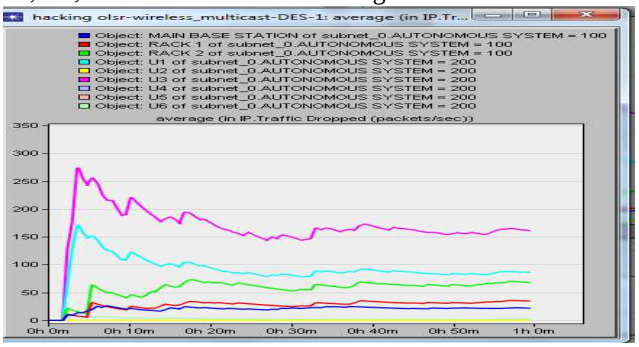


Fig .33 IP-Traffic dropped (pkts/sec)

From fig 33 the higher is UAV3 & the lower is UAV4 and numerically the UAV3 average is 160.72pkts/sec, UAV4 average is 0.02 pkt/sec.

c. TORA scenario under attack using DOS attack

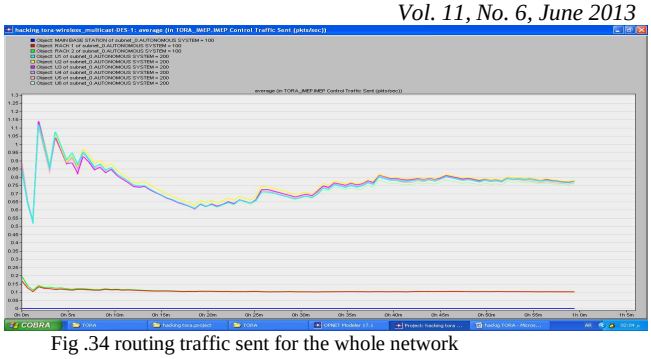


Fig .34 routing traffic sent for the whole network

From fig.34 the higher is UAV2 & the lower is Rack 1 in routing traffic sent & numerically the UAV2 average is 0.77944 pkts/sec , Rack1 average is 0.10083 pkts/sec.

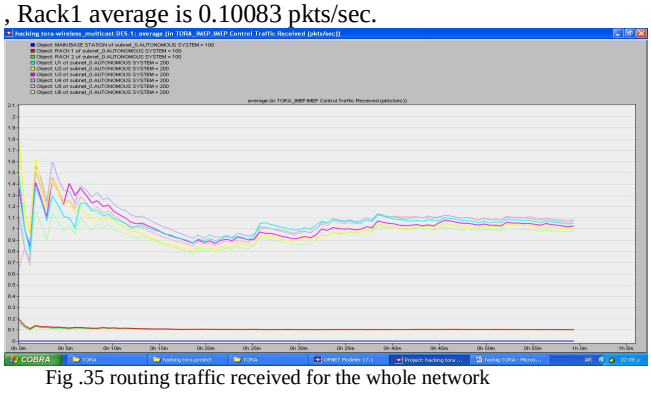


Fig .35 routing traffic received for the whole network

From fig.35 the higher is UAV 5 & the lower is rack1, 2 in routing traffic received & numerically UAV 5 average is 1.0858 pkts/sec , rack 1 average is 0.1014 pkts/sec , rack2 average is 0.1008 pkts/sec.

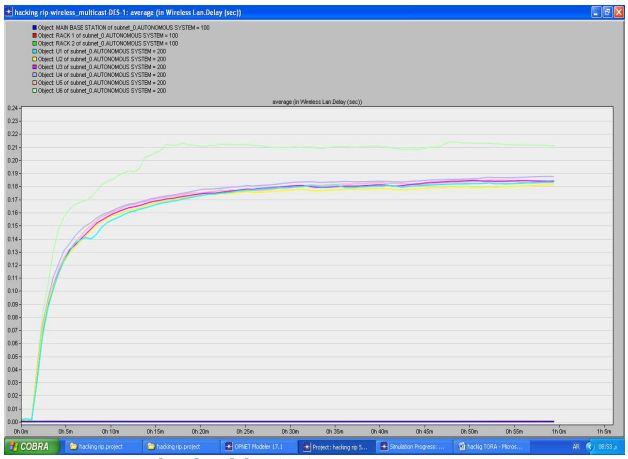


Fig .36 wireless lan delay (sec)

From fig.36 the higher is UAV 6 & the lower is the base station in wireless lan delay & numerically UAV 6 average is 0.21119sec, base station average is 0.00005 sec.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

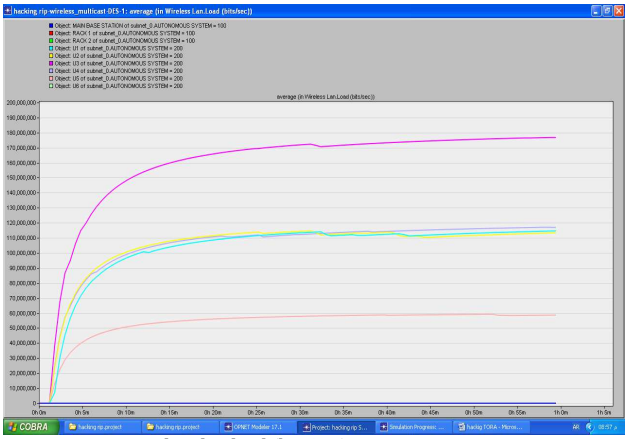


Fig .37 wireless lan load (bits/sec)

From fig.37 the higher is UAV 3 & the lower is Rack 1 in wireless lan load & numerically UAV 3 average is 176.804.286 bits/sec, Rack 1 average is 64 bits/sec.

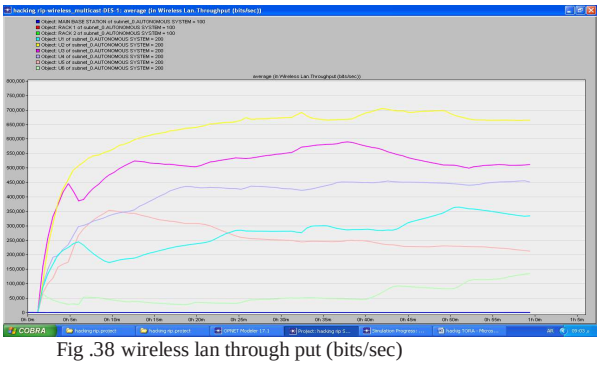


Fig .38 wireless lan through put (bits/sec)

From fig .38 the higher is UAV2& the lower is the base station and numerically the UAV2 average is 664.130 bits/sec & the base station average is 17 bits/sec.

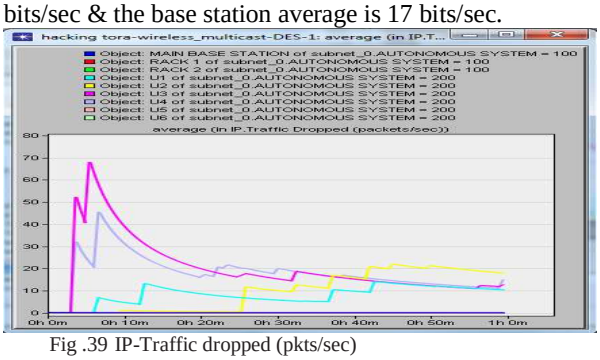


Fig .39 IP-Traffic dropped (pkts/sec)

From fig 39 the higher is UAV2 & the lower is Rack1 and numerically the UAV2 average is 17.950pkts/sec, Rack1 average is 0.002 pkt/sec.

iv. SECURING ROUTING PROTOCOLS (AODV, OLSR, TORA)

Security plays a more important part in Wireless Communication Systems and espicially in UAVs . This is mainly because of the ubiquitous nature of the Wireless Medium that makes it more susceptible to Security Attacks. In the Wireless Medium, anyone can listen to whatever is being sent over the network but in UAVs it is Unacceptable. To make things worse, any tapping or eaves-dropping cannot even be detected in a medium as ubiquitious as the Wireless Medium. Thus Security plays a vital role for the successful operation of UAVs. The only way to prevent such unauthorized use of communication resources and to protect user privacy, is the use cryptographic techniques to provide security, authentication and access-control[21]. The terms authentication and privacy are generally related to each other, since the derivation of the session key for further encryption of user data is done at the Authentication stage. In providing a secure networking environment some or all of the following service may be required

1. Authentication: This service verifies the identity of node or a user, and to be able to prevent impersonation. In wired networks and infrastructure-based wireless networks, it is possible to implement a central authority at a point such as a router, base station, or access point. But there is no central authority in UAV, and it is much more difficult to authenticate an entity. Authentication can be providing using encryption along with cryptographic hash function, digital signature and certificates.

2. Confidentiality: Keep the information sent unreadable to unauthorized users or nodes. UAV uses an open medium, so usually all nodes within the direct transmission range can obtain the data. One way to keep information confidential is to encrypt the data, and another technique is to use directional antennas. It also ensures that the transmitted data can only be accessed by the intended receivers.

3. Integrity: Ensure that the data has been not altered during transmission. The integrity service can be provided using cryptography hash function along with some form of encryption. When dealing with network security the integrity service is often provided implicitly by the authentication service.

4. Availability: Ensure that the intended network security services listed above are available to the intended parties when required. The availability is typically endure by redundancy, physical protection and other non-cryptographic means, e.g. use of robust protocol.

5. Non-repudiation: Ensure that parties can prove the transmission or reception of information by another party, i.e. a party cannot falsely deny having received or sent certain data. By producing a signature for the message, the entity cannot later deny the message. In public key cryptography, a node A signs the message using its private key. All other nodes can verify the signed message by using A's public key, and A cannot deny that its signature is attached to the message.

6. Access Control: To prevent unauthorized use of network services and system resources. Obviously, access control is tied to authentication attributes. In general, access control is the most commonly thought of service in both network communications and individual computer systems[22].

So by applying encryption & authentication algorithms. The three scenarios are secured by implementing the following:

1) Authentication:

a) Authentication algorithm : MD5

b) Authentication method : RSA signature

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

2) *Encryption:*
   a) *3DES*
   b) *IP-sec proposals/transform sets*
   c) *Global properties:*
      - Protocol : Bundle (AH+ESP)
      - Authentication algorithm : HMAC MD5
      - Encryption algorithm : 3DES
   d) *Security Association/static crypto map sets:-*
      - Type : IKE
      - Mode: Transport mode
      - Direction : Bidirectional

These methods are easy to be used using Opnet simulator and this is one of the reason we use Opnet Modeler v17.

A. *Securing AODV scenario*

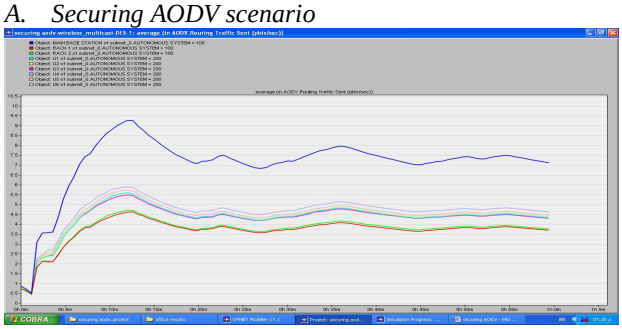


Fig .40 routing traffic sent for the whole network

From fig.40 the higher is the base station & the lower is Rack 1 in routing traffic sent & numerically the base station average is 7.1150 pkts/sec , Rack1 average is 3.6964 pkts/sec.

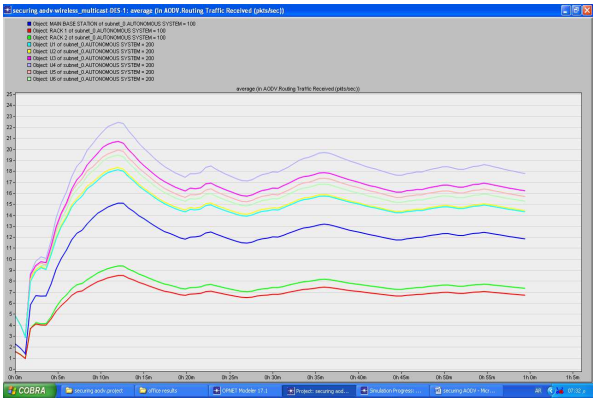


Fig .41 routing traffic received for the whole network

From fig.41 the higher is UAV 4 & the lower is rack1, 2 in routing traffic received & numerically UAV 4 average is 17.772 pkts/sec , rack 2 average is 7.339 pkts/sec , rack1 average is 6.707 pkts/sec.

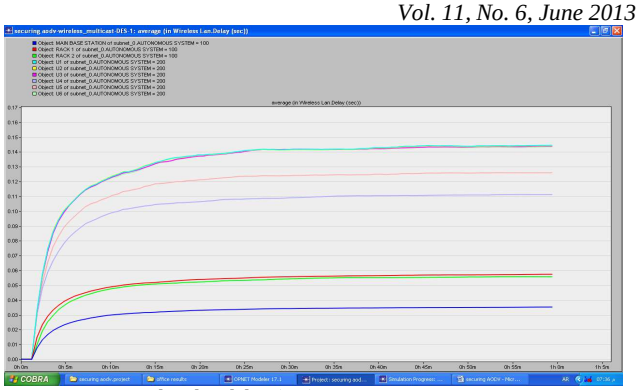


Fig .42 wireless lan delay (sec)

From fig.42the higher is UAV 1 & the lower is the base station in wireless lan delay & numerically UAV 1 average is 0.14457 sec, base station average is 0.03534 sec.

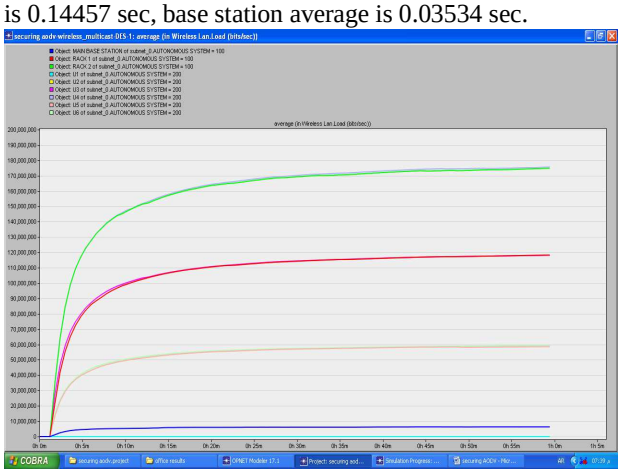


Fig .43 wireless lan load (bits/sec)

From fig.43the higher is UAV 4 & the lower is UAV1 in wireless lan load & numerically UAV 4 average is 175.662.321 bits/sec, UAV1 average is 3.548 bits/sec.

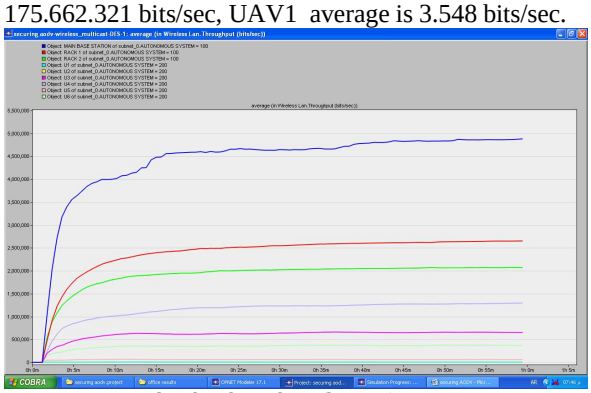


Fig .44 wireless lan through put(bits/sec)

From fig .44 the higher is the base station& the lower is UAV1 and numerically the base station average is 4,881,895 bits/sec & UAV1 average is 6.215bits/sec.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

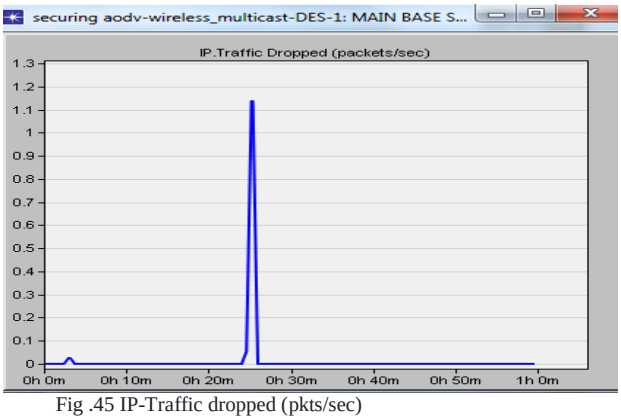


Fig .45 IP-Traffic dropped (pkts/sec)

From fig .45only the base station has packet dropped and numerically it is 0.012222 pkt/sec.

B. Securing OLSR scenario

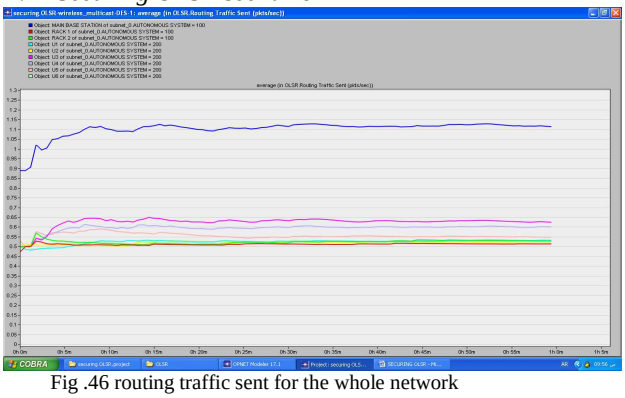


Fig .46 routing traffic sent for the whole network

From fig.46the higher is the base station & the lower is Rack 1 in routing traffic sent & numerically the base station average is 1.1136 pkts/sec , Rack1 average is 0.5133 pkts/sec.

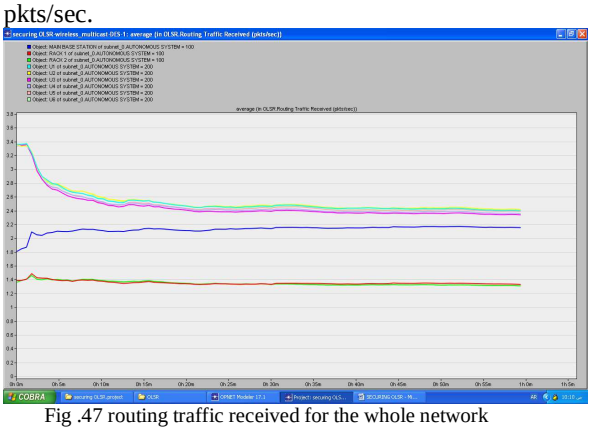


Fig .47 routing traffic received for the whole network

From fig.47the higher is UAV 2 & the lower is rack2, in routing traffic received & numerically UAV 2 average is 2.4158 pkts/sec , rack 2 average is 1.3094 pkts/sec .

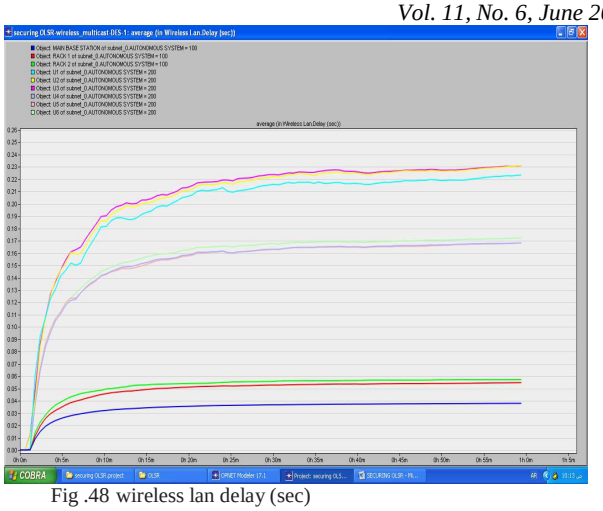


Fig .48 wireless lan delay (sec)

From fig.48the higher is UAV 2 & the lower is the base station in wireless lan delay & numerically UAV 2 average is 0.23117 sec, base station average is 0.03802 sec.

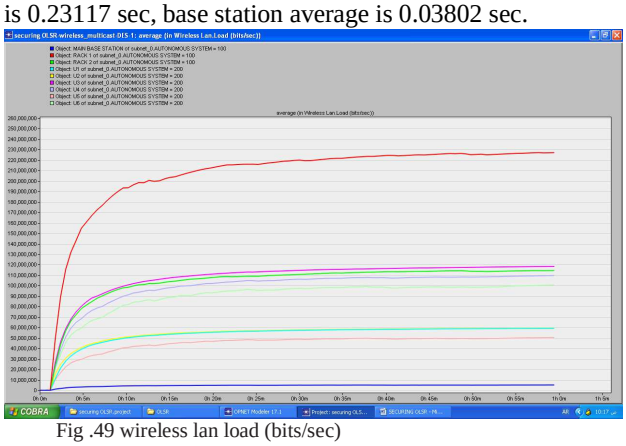


Fig .49 wireless lan load (bits/sec)

From fig.49 the higher is Rack1 & the lower is the base station in wireless lan load & numerically Rack1 average is 227.140.469 bits/sec, base station average is 5.036.449 bits/sec.

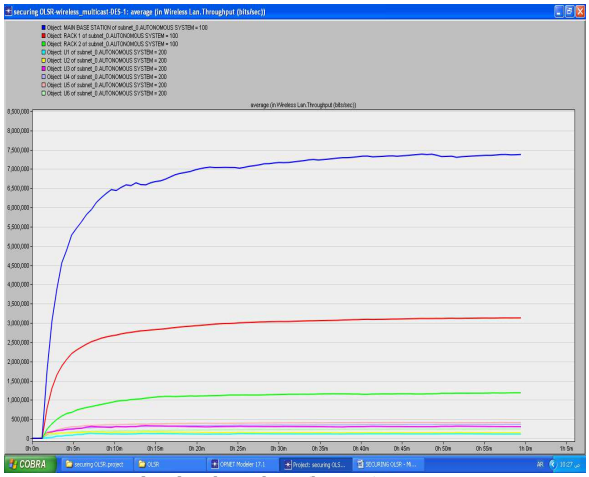


Fig .50 wireless lan through put(bits/sec)

From fig .50 the higher is the base station & the lower is UAV1 & numerically the base station average is 7,377,636 bits/sec, UAV1 average is 112.726 bits/sec

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

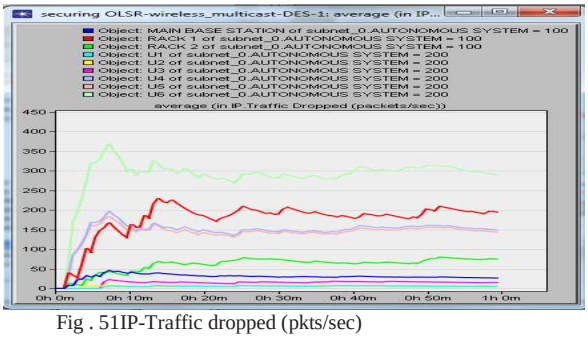


Fig . 51IP-Traffic dropped (pkts/sec)

From fig .51the higher is UAV6 & the lower is UAV2 and numerically UAV6 average is 288.89pkts/sec , UAV2 average is 4.94 pkts/sec.

C. Securing TORA scenario

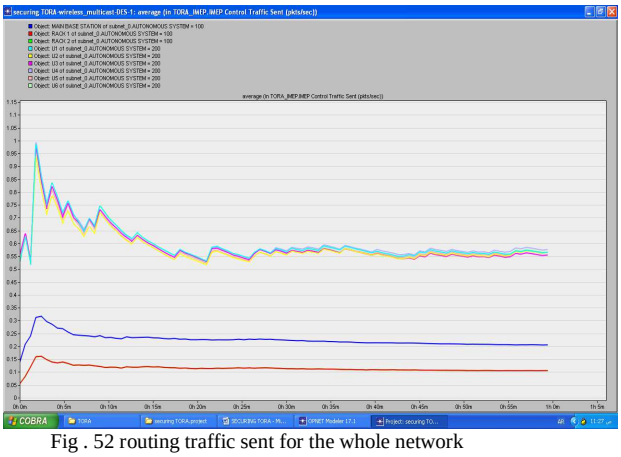


Fig . 52 routing traffic sent for the whole network

From fig.52the higher is UAV4 & the lower is Rack 1 in routing traffic sent & numerically UAV4 average is 0.57639 pkts/sec , Rack1 average is 0.10556 pkts/sec.

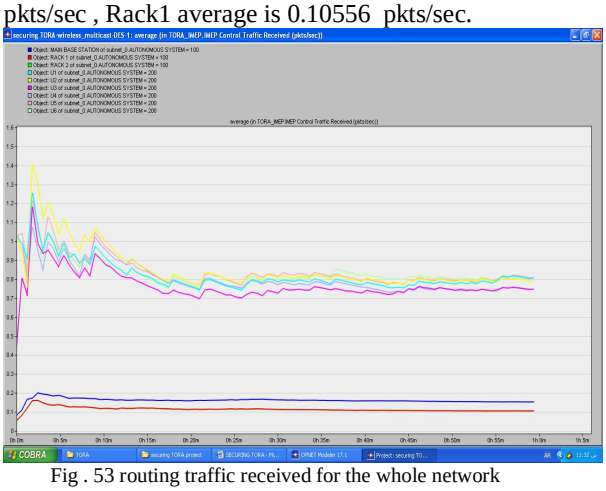


Fig . 53 routing traffic received for the whole network

From fig.53the higher is UAV 5 & the lower is rack1 in routing traffic received & numerically UAV 5 average is 0.80889 pkts/sec , rack 1 average is 0.10556 pkts/sec .

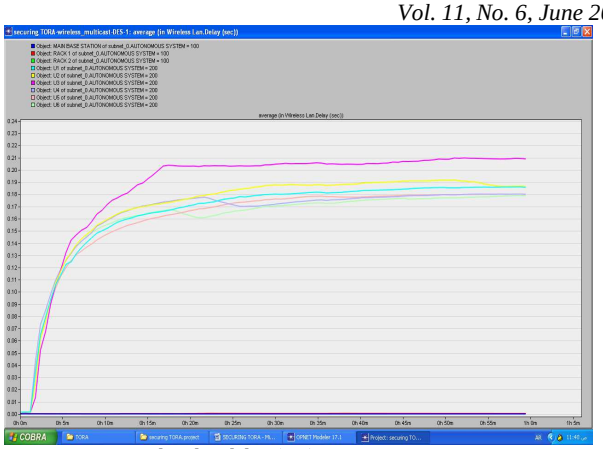


Fig .54 wireless lan delay (sec)

From fig.54the higher is UAV 3 & the lower is the base station in wireless lan delay & numerically UAV 3 average is 0.20901 sec, base station average is 0.00005 sec.

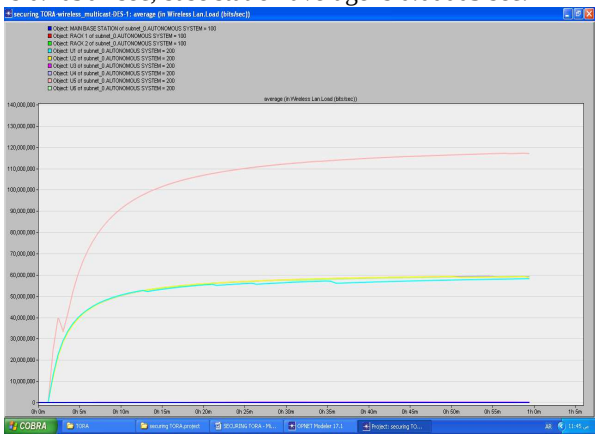


Fig .55 wireless lan load (bits/sec)

From fig.55the higher is UAV5 & the lower is Rack1 in wireless lan load & numerically UAV5 average is 117.016.313 bits/sec, Rack1 average is 170 bits/sec.

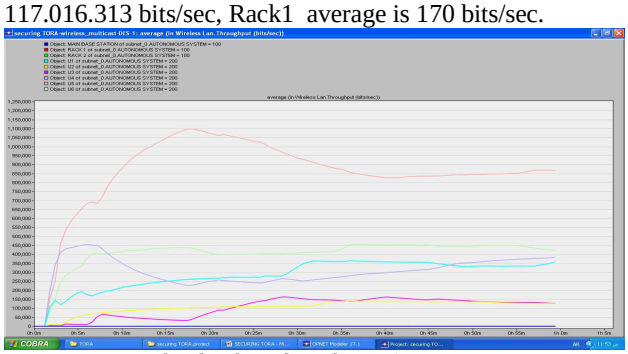


Fig. 56 wireless lan through put(bits/sec)

From fig .56the higher is UAV5 & the lower is the base station and numerically UAV5 average is 866.399 bits/sec , the base station average is 17 bits/sec.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

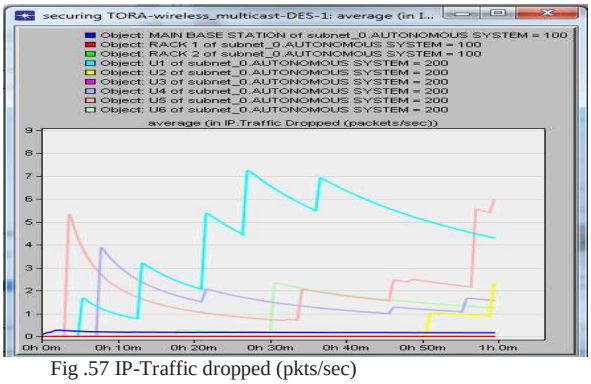


Fig .57 IP-Traffic dropped (pkts/sec)

From fig .57the higher is UAV5 with average of 5.9686pkts/sec.

v. Simulation Analysis

In order to compare the results taken we choose one node only (UAV3) for simplicity to highlight our results.

- Table .1 shows a comparison for AODV protocol in all the scenarios.

Point of compareson	Normal	Under attack	Securing
Traffic sent(pkts/sec)	8.981	4.6711	4.3025
Traffic recevied(pkts/sec)	21.770	15.981	16.207
Delay(sec)	0.34951	0.20996	0.14401
Load(bits/sec)	56,296,636	175,361,935	118,164,456
Through put(bits/sec)	79.314	738.725	652.524
Traffic dropped(pkt/sec)	-	-	-

Table. 1 AODV in all scenarios

From table .1 we notice that AODV routing protocol has no traffic dropped in all the scenarios but the load under the attack was too high and this may cause increase in the collision of packets which may cause data dropped, it also may cause wastage of channel.

- Table .2 shows a comparison for OLSR routing protocol in all scenarios.

Point of compareson	Normal	Under attack	Securing
Traffic sent(pkts/sec)	0.6236	0.6092	0.6242
Traffic recevied(pkts/sec)	2.4403	2.5319	2.3389
Delay(sec)	0.18473	0.09585	0.23101
Load(bits/sec)	178,482,209	228,149,644	118,343,123
Through put(bits/sec)	606.780	2,027,877	304.524
Traffic dropped(pkt/sec)	7.74	160.72	14.72

Table. 2 OLSR in all scenarios

From table .2 we notice that OLSR routing protocol traffic dropped is high under attack than the other two protocols as shown in table 1, 2 &3 but when applying the security mechanisms used it is reduced but still also the higher traffic dropped as compared to AODV or TORA.

Also the load under attack is too high if it is compared to the other protocols or even if it is compared to the normal scenario of OLSR and when applying the security mechanisms used it is reduced but still also the higher if it is compared to the other two protocols.

- Table .3 shows a comparison for TORA protocol in all the scenarios.

Point of compareson	Normal	Under attack	Securing
Traffic sent(pkts/sec)	0.32611	0.77722	0.55500
Traffic recevied(pkts/sec)	0.54000	1.0258	0.74722
Delay(sec)	0.00033366	0.18413	0.20901
Load(bits/sec)	29.801	176,804,288	129.394
Through put(bits/sec)	29.838	511.446	127.375
Traffic dropped(pkt/sec)	-	12.864	-

Table. 3 TORA in all scenarios

From table .3 we notice that TORA routing protocol under attack has traffic dropped but still lower than OLSR if it is compared to and when applying the security mechanisms used we found that there is no traffic dropped.

- Table .4 shows a comparison for SAODV protocol &the secured protocols in all the scenarios.

Performance parameters	SAODV	Secured AODV	Secured OLSR	Secured TORA
Type	Reactive	Reactive	Proactive	Reactive
MANET Protocol	AODV	AODV	OLSR	TORA
Encryption	Asym	Sym	Sym	Sym
Synchronization	No	Yes	Yes	Yes
Trust Authority	CA	CA	CA	CA
Authentication	Yes	Yes	Yes	Yes
Confidentiality	No	Yes	No	Yes
Integrity	Yes	Yes	No	Yes
Non-repudiation	Yes	No	No	No
Anti-Spoofing	Yes	Yes	No	Yes
Dos Attacks	No	No	Yes	No

Table. 4 Comparison between SAODV & secured protocols

Our issue here is to make a comparison between a real secured protocol which is implemented in the table by SAODV and the protocols we tried to secure and from the above comparison it is found that the ordinary protocols we tried to secure gives a results quite close to the results given by SAODV.

vi. Conclusion

From the results taken in section 5 we found that using AODV , TORA routing protocols (reactive protocols) with secured mechanisms are more efficient and secured than using OLSR routing protocol. This result makes as quite close to the results taken by [15,16].The conclusion of this paper on the basis of the results is that the AODV is better in those scenarios & it is followed by TORA in the same conditions. Also we found that when applying encryption & authentication algorithms we get better results.

References

[1]. DARPA Intrusion Detection Evaluation project, at http://www.ll.mit.edu/IST/ideval/data/1999/1999_data_index.html.

[2]. Shervin Ehrampoosh and Ali Khayatzadeh Mahani," Secure Routing Protocols: Affections on MANETs Performance,"1st international conference on communication engineering 22-24 December 2010

[3]. C. E. Perkins, E. M. Royer,"Ad-Hoc On-Demand Distance Vector Routing", Proc. 2nd IEEE Wksp. Mobile comp. Sys. And Apps. Feb,1999, pp. 90-100.

[4]. C.E. Perkins, E. M. Belding-Royer, S. R. Das. "Ad-Hoc On-Demand Distance Vector Routing", IETF RFC 3561, July2003.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

[5]. C. Sreedhar, Dr. S. Madhusudhana Verma, Dr. N. Kasiviswanath“Performance Analysis of Secure Routing Protocols in Mobile Ad-Hoc Networks”, IJCST Vol. 3, Issue 1, Jan. - March 2012.
[6]. Junaid Arshad, Mohammad Ajmal Azad,“Performance Evaluation of Secure on-Demand Routing Protocols for Mobile Ad-Hoc Networks”, 2006 IEEE.
[7]. M. F. Juwad, H. S. Al Raweshidy,“Experimental Performance Comparisons between SAODV & AODV”, IEEE Second Asia International Conference on Modelling & Simulation, 2008.
[8].P.Jacquet, P.Muhlethaler, T.Clausen, A.Laouiti, A.Qayyum, L.Viennot “ Optimized Link State Protocol for Ad Hoc Networks”, Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International,30-30 Dec. 2001
[9]. A.Qayyum, L. Viennot, and A.Laouiti. Multipoint relaying: An efficient technique for flooding in mobile wireless networks. Technical Report RR-3898, INRIA,February 2000.
[10]. Amir Qayyum. Wireless Networks: Hiperlan. Master’s thesis, Universite de Paris-Sud, Orsay, France, September 1996.
[11].Pankaj Palta, Sonia Goyal,” Comparison of OLSR and TORA Routing Protocols Using OPNET Modeler, International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 5, July – 2012.
[12]. Aleksandr Huhtonen, "Comparing AODVand OLSR Routing Protocols", Seminar on Internetworking 2004.
[13]. Monika Rajput, Pallavi Khatri, Alankar Shastri and Keshav Solanki ,"Comparison ofAd-Hoc Reactive Routing Protocols using OPNET Modeler" IEEE 2010.
[14]. Algorithms and Protocols for wireless and mobile Ad Hoc Networks by AZZEDINE BOUKERCHE.
[15].Anuj K. Gupta,Dr.Harsh Sadawarti,Dr.Anil K. Verma"Performance analysis of AODV,DSR&TORA Routing protocols"international journal of Engineering and Technology,Vol.2,No.2,April 2010.
[16]. C. Sreedhar, Dr. S. Madhusudhana Verma, Dr. N. Kasiviswanath"Performance Analysis of Secure Routing Protocols inMobile Ad-Hoc Networks",IJCST Vol. 3, Issue 1, Jan. - March 2012.
[17]. Chris Karlof, David Wagner"Secure routing in wireless sensor networks:attacks &counter measures",Adhoc networks1(2003) 293-315.
[18].Shio kuman singh,Mp singh and Dk singh "Routing protocols in wireless sensor networks-Asurvey,international journal of computer science &engineering survey(IJCSES) Vol.1,No.2,November 2010.
[19].Panagiotis papadimitratos and Zygmunt J.Haas"Secure routing for mobile adhoc networks" ,In proceedings of the scs communication networks & distributed systems Modeling & Simulation conference (CNDS2002),San Antonio , TX, January 27-31,2002.
[20].Ahmed A.Radwan,Tarek M.Mahmoud and Essam H.Houssein "Performance Measurement of some Mobile Adhoc network routing protocols ", IJCSI international journal of computer science issues, vol8,issue1, January 2011.
[21] VIJAYA CHANDRAN RAMASAMI, KUID 698659 "SECURITY, AUTHENTICATION AND ACCESS CONTROL FOR MOBILE COMMUNICATIONS",EECS Department, The University of Kansas .
[22] Amol Bhosle, Yogadhar Pandey"Applying Security to Data Using Symmetric Encryption in MANET",International Journal of Emerging Technology and Advanced Engineering Volume 3, Issue 1, January 2013.

Survey on Internet-based Mobile Ad Hoc Networking

Omung Goyal
Dept. of CSE
UIT RGPV,Bhopal

Swati Jaiswal
Dept. of CSE
UIT RGPV,Bhopal

Prateek Poste
Dept. of ECE
UIT RGPV,Bhopal

Abstract: Internet-based Mobile Ad Hoc Networking (MANET) is an emerging technology that supports self-organizing mobile networking infrastructures. This is expected to be of great use in commercial applications for the next generation internet users. A number of technical challenges are faced today due to the heterogeneous, dynamic nature of this hybrid MANET. A new hybrid routing scheme AODV ALMA is proposed, which act simultaneously combining mobile agents to find path to the gateway and on-demand distance vector approach to find path in local MANET is one of the unique solution. An adaptive gateway discovery mechanism based on mobile agents making use of pheromone value, pheromone decay time and balance index is used to estimate the path and next hop to the gateway. The mobile nodes automatically configure the address using mobile agents first selecting the gateway and then using the gateway prefix address. The mobile agents are also used to track changes in topology enabling high network connectivity with reduced delay in packet transmission to Internet.
Keywords: Hybrid MANET, mobile agents, AODV ALMA, adaptive gateway discovery.

Introduction

A mobile ad hoc network (MANET) is an autonomous network that can be formed without need of any established infrastructure or centralized administration. It normally consists of mobile nodes, equipped with a wireless interface, that communicate with each other. Because these kinds of networks are very spontaneous and self-organizing, they are expected to be very useful. It is also highly likely that a user of the network will have the need to connect to the Internet. The Internet Engineering Task Force (IETF) has proposed several routing proto-cols for MANETs, such as Ad hoc On-Demand Distance Vector (AODV) [1], Dynamic Source Routing (DSR) [2], Optimized Link State Routing Protocol (OLSR) [3] and Topology Dissemination Based on Reverse-Path Forwarding (TBRPF) [4]. However, these protocols were designed for communication within an autonomous MANET, so a routing protocol needs to be modified in order to achieve routing between a mobile device in a MANET and a host device in a wired network (e.g. the Internet). To achieve this network interconnection, gateways that understand not only the IP suite, but also the MANET protocol stack, are needed. Thus, a gateway acts as a bridge between a MANET and the Internet and all communication between the two networks must pass through any of the gateways.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Researchers presently make use of the social behaviors of insects and of other animals for developing a variety of different organized behaviors at the system level. In particular, ants have inspired a number of methods and techniques among which the most studied and the most successful is the general purpose optimization technique know as Ant Colony Optimization (ACO). ACO takes inspiration from the
For aging behaviour of some ant species. These ants deposit pheromone on the ground in order to mark some favourable path that should be followed by other members of the colony. Because of the robustness and efficiency of ACO they have recently become a source of inspiration for the design of routing algorithms. Following the ACO principles, Ant like Mobile agents ALMA) are simple packets that explore the network and collect data for routing [2, 3]. They communicate with one another indirectly by exchanging routing information by writing and reading data to from routing tables. This is similar to ants exchanging food information with each other. By using mobile agent exploration, the idea is to put the intelligence across the network and make the routing distributed with high connectivity. The important constraint in using the mobile agents for routing is that the overhead is proportional to the number of agents present in the network. This in return reduces the overhead compared with traditional routing protocols that use frequent updates and reactive updates. Figure 1 shows the architecture in which the mobile nodes form local MANET access with multi-hop wireless links to provide service within them. The Internet Gateway (IGW) nodes form global access to provide internet service to mobile nodes. The main technical challenges faced today in this scenario are Internet gateway discovery, address auto-configuration and reaching the destination. The ACO mechanism is used as an adaptive reactive gateway discovery approach integrated with address auto- configuration. The new hybrid scheme, proposed in this paper, combines local access with global access, named as AODV ALMA. The remainder of the paper is organized as follows. The paper gives an overview of the related work in section II. The new protocol description is given in section III. The performance analysis is simulated and compared in section IV. The future enhancement is given as conclusion in section V.

II. Related Existing Proposals

The first proposal by Broch et al. [2] is based on integration of Mobile IP and MANET, employ a source routing protocol. The second proposal MIPMANET [6] followed a similar approach based on AODV, but works only with Mobile IPv4 approach, because it requires Foreign Agents (FAs). The proposal by Ammari et al. [1] gives a mobile gateway approach based on DSDV protocol and analyzed the performance. In Ratanchandani et al. [8] proposal, a hybrid scheme for propagating the advertisements up to a certain number of hops is given. The mobile nodes located out of this propagation scope will reactively find the gateways when needed. The performance of this approach depends on the Time-To-Live (TTL) value, set for particular scenario and network conditions under consideration. In Lee [5] proposal, another more sophisticated hybrid approach the advertisements are sent out only when changes in topology are detected. However, they rely on a source based routing protocol, which limits the applicability to particular type of routing protocol. In Ruiz et al. [9] proposal, an adaptive gateway discovery mechanism that outperforms existing hybrid approaches is proposed. The key scheme is that TTL value for proactive gateway advertisements is adjusted dynamically to network conditions. The mobile nodes out of the propagation scope reactively find the gateways. The proposals available till date focused only on gateway discovery and are not a complete solution for inter-working with fixed IP networks. The complete proposals available to integrate both auto configuration and gateway discovery

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

are namely Wakikawa [12], Jelger [4] and Singh et al. [11]. In Wakikawa [12], mobile nodes discover the gateway either reactively or proactively to configure their IP address automatically. In the proactive approach, gateway periodically floods the gateway advertisement messages up to certain number of hops in the MANET. In the reactive approach, mobile nodes flood the gateway solicitation message to all gateways reachable. The gateway receiving the gateway solicitation will unicast back the gateway advertisement to the mobile nodes. The gateway advertisement message contains the global IPv6 address of the gateway, the network prefix advertised by the gateway, the prefix length and the lifetime associated with this information.

III. Proactive Gateway Discovery

The proactive gateway discovery is initiated by the gateway itself. The gateway periodically broadcasts a gateway advertisement (GWADV) message with the period determined by ADVERTISEMENT_INTERVAL [2]. The advertisement period must be chosen with care so that the network is not flooded unnecessarily. The mobile nodes that receive the advertisement, create a (or update the) route entry for the gateway and then rebroadcast the message. To assure that all mobile nodes within the MANET receive the advertisement, the number of retransmissions is determined by NET_DIAMETER defined by AODV. However, this will lead to enormously many unnecessary duplicated advertisements. A conceivable solution that prevents duplicated advertisements is to introduce a “GWADV ID” field in the advertisement message format similar to the “RREQ ID” field in the RREQ message format (see Sect. The advantage of this approach is that there is a chance for the mobile node to initiate a handover before it looses its Internet connection. The disadvantage is that since a control message is flooded through the whole MANET periodically, limited resources in a MANET, such as power and bandwidth, will be used a lot.

Reactive Gateway Discovery

The reactive gateway discovery is initiated by a mobile node that is to create or update a route to a gateway. The mobile node broadcasts a RREQ with an ’I’ flag (RREQ_I) to the ALL_MANET_GW_MULTICAST [6] address, i.e. the IP address for the group of all gateways in a MANET. Thus, only the gateways are addressed by this message and only they process it. Intermediate mobile nodes that receive a RREQ_I are not allowed to answer it, so they just rebroadcast it. When a gateway receives a RREQ_I, it unicasts back a RREP_I which, among other things, contains the IP address of the gateway. The advantage of this approach is that control messages are generated only when a mobile node needs information about reachable gateways. The disadvantage of reactive gateway discovery is that a handover cannot be initiated before a mobile node looses its Internet connection.

Hybrid Gateway Discovery

To minimize the disadvantages of the proactive and reactive strategies, they can be combined into a hybrid proactive/reactive method for gateway discovery. For mobile nodes in a certain range around a gateway, proactive gateway discovery is used while mobile nodes residing outside this range use reactive gateway discovery to obtain information about the gateway. The gateway periodically broadcasts a GWADV message. Upon receipt of the message, the mobile nodes update their routing

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

table and then rebroadcast the message. The maximum number of hops a GWADV can move through the MANET is determined by ADVERTISEMENT_ZONE. This value defines the range within which proactive gateway discovery is used. When a mobile node residing outside this range needs gateway information, it broadcasts a RREQ_I to the ALL_MANET_GW_MULTICAST address. Mobile nodes receiving the RREQ_I just rebroadcast it. When a gateway receives a RREQ_I, it sends a RREP_I towards the source.

Protocol Description: AODV

ALMA

In hybrid MANET network, connectivity of mobile nodes with gateway nodes and mobile nodes with other active mobile nodes is uncertain due to the dynamic network topology. The mobility of mobile nodes cause more delay in finding the route to destination either in Internet or in local MANET. The newly proposed hybrid mechanism AODV ALMA overcomes this by combining the on demand distance vector approach with ACO. The on demand distance vector approach, i.e., AODV used to find the path in local MANET. The ACO approach, i.e., ALMA (Ant like Mobile Agents) used to find the path to the gateway. The hybrid mechanism operates simultaneously when a route to the destination is needed by the source mobile node.

Gateway Discovery in AODV ALMA

The proposed protocol is used to discover paths and to maintain paths in local MANET for dynamic topology changes. The mobile nodes initially use the available temporary address for broadcasting before configuring the address based on the gateway. The mobile nodes flood ACO messages and AODV messages simultaneously to discover routes either in global Internet or in local MANET. The ACO message first sent to discover the gateway is the Forward Routing Mobile Agent (FRMA) message. The FRMA message is multicast with the all manet gateway multicast address. The intermediate mobile nodes in the path to the gateway receive the FRMA message and mark the message with its node information into the history field. The FRMA message on traversing the link through the intermediate nodes to the gateway will get the link stability factor in terms of the pheromone value and pheromone decay time. The intermediate node after marking the FRMA message will rebroadcast in local MANET until it discovers the gateway. The FRMA message keeps the list of mobile nodes [1, 2, …, n] it has visited in the history field. The Route Request Message (RREQ) of AODV is used to find the path in local MANET. The second stage in the proposed protocol is route maintenance in local MANET and to Internet. In local MANET it is responsible to keep paths to local destination by getting special packets i.e. hello messages from intermediate nodes. In Internet it is responsible to keep paths to the gateway through intermediate nodes for high network connectivity. The route maintenance to gateway for getting services from Internet is done automatically using the data packets forwarded through the intermediate active nodes selected in gateway discovery. The mobile agents traversing the link establish a pheromone track by setting the pheromone values from the source node to the gateway node. The path established keeps updating the pheromone value in association with the decay factor, i.e., pheromone decay time. The pheromone decay time is set based on the link life time predicated by mobile agents. The pheromone value gets increased by a predefined value to maintain the route based on the sensitivity

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

factor F for each data packet traversed. The pheromone value gets decreased by a predefined value to know the path is not used any more.

IV. Performance Evaluation

The performance evaluation of proposed protocol compared with existing proposals of Wakikawa and Jelger using the network simulator ns-2. The source code of AODV protocol modified with the proposal of Wakikawa as AODV Wahiawa and OLSR protocol modified with the proposal of Jelger as OLSR Jelger. The proposed protocol AODV ALMA uses mobile agent code available as open source. The simulation scenario description is given in Table 3 with the traffic type set for all nodes as same ftp traffic.

Packet Delivery Ratio (PDR)

Transmission range 250 mts
Physical layer two-ray propagation
Link layer IEEE 802.11
Queue buffer size 64 packets
Routing buffer size 50 packets
Channel capacity 2 Mbps
No. of gateway 3
No. of mobile node 50
Area 1500 x 300 mts

The PDR of AODV ALMA is giving high throughput with that of the reactive AODV Wakikawa. The change of performance for high pause times caused when the network is quite sparse and static giving a condition that nodes become unreachable for long periods. The OLSR Jelger has very less PDR compared to the proposed protocol. The reason for this in OLSR Jelger is that link break in selected path to either gateway or local MANET handled after some time gap. In AODV ALMA the link break to the gateway is monitored by the change in the pheromone value. The pheromone decay time associated with life time of the link. The route maintenance is done with the threshold value for each link in the path to the gateway. After carefully analyzing the simulation results, AODV ALMA use the pheromone value to adaptively select next hop to the gateway giving topology changes strong to mobility.

V. Conclusion

This paper analyses a new hybrid adaptive gateway discovery mechanism for hybrid MANET using mobile agents combined with reactive distance vector. The performance tradeoffs and limitations with existing proposals are compared with the proposed protocol using simulation. The proposed protocol gives high network connectivity to gateway and reduced end-to-end delay in packet transmission to destination. The results compared for mobility condition only, it can also be tested for different traffic and scalability conditions. The proposed protocol can be extended with other internetworking related issues such as improved DAD mechanism, efficient support of DNS, discovery of application and network services, network authentication and integrated security mechanism.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

References

[1] Ammari H. and El-Rewini H., "Performance Evaluation of Hybrid Environments with Mobile Gateways," in Proceedings of the 9th International Sympesium on Computing and Communications, vol. 1, no. 1, pp. 152-57, June 2004.

[2] Broch J., Maltz D., and Johnson D., "Supporting Hierarchy and Heterogeneous Interfaces in Multi-Hop Wireless Ad Hoc Networks," in Proceedings of the Workshop on Mobile Computing, Australia, pp. 85-97, June 1999.

[3] Choudhary R., Bhandhopadhyay S., and Paul K., "A Distributed Mechanism for Topology Discovery in Ad Hoc Wireless Networks Using Mobile Agents," in Proceedings of Mobicom, pp. 145-146, August 2000.

[4] Jelger C., Noel T., and Frey A., "Gateway and Address Auto Configuration for IPv6 Ad Hoc Networks," Internet-draft, April 2004.

[5] Lee J., "Hybrid Gateway Advertisement Scheme for Connecting Mobile Ad Hoc Networks to the Internet," in Proceedings of the 57th IEEE VTC 2003, vol. 1, no. 1, pp. 191-95, April 2003.

[6] Lei H. and Perkins C., "Ad Hoc Networking with Mobile IP," in Proceedings of the Second European Personal Mobile Communications Conference, Germany, pp. 197-210, 1997.

[7] Ratanchandani P. and Kravets R., "A Hybrid Approach to Internet Connectivity for Mobile Ad Hoc Networks," in Proceedings of IEEE Wireless Communications and Networking Conference, pp. 295-304, March 2003.

[8] Ruiz P. and Gomez-Skarmeta A., "Adaptive Gateway Discovery Mechanisms to Enhance Internet Connectivity for Mobile Ad Hoc Networks," Ad Hoc and Sensor Wireless Networks, vol. 1, pp. 159-77, March 2005.

[9] Scott M., Joseph P., and Gregory H., "Internet Based Mobile Ad Hoc Networking," IEEE Internet Computing, vol. 3, no. 4, pp. 63-70, July-August 1999.

[10] Singh S., Perkins C., Ruiz P., and Clausen T., "Ad Hoc Network Auto Configuration: Definition and Problem Statement," http://tools.ietf.org/id/draft-singh-autoconf-adp-00.txt, July 2005.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

IJCSIS REVIEWERS' LIST

Assist Prof (Dr.) M. Emre Celebi, Louisiana State University in Shreveport, USA
Dr. Lam Hong Lee, Universiti Tunku Abdul Rahman, Malaysia
Dr. Shimon K. Modi, Director of Research BSPA Labs, Purdue University, USA
Dr. Jianguo Ding, Norwegian University of Science and Technology (NTNU), Norway
Assoc. Prof. N. Jaisankar, VIT University, Vellore,Tamilnadu, India
Dr. Amogh Kavimandan, The Mathworks Inc., USA
Dr. Ramasamy Mariappan, Vinayaka Missions University, India
Dr. Yong Li, School of Electronic and Information Engineering, Beijing Jiaotong University, P.R. China
Assist. Prof. Sugam Sharma, NIET, India / Iowa State University, USA
Dr. Jorge A. Ruiz-Vanoye, Universidad Autónoma del Estado de Morelos, Mexico
Dr. Neeraj Kumar, SMVD University, Katra (J&K), India
Dr Genge Bela, "Petru Maior" University of Targu Mures, Romania
Dr. Junjie Peng, Shanghai University, P. R. China
Dr. Ilhem LENGLIZ, HANA Group - CRISTAL Laboratory, Tunisia
Prof. Dr. Durgesh Kumar Mishra, Acropolis Institute of Technology and Research, Indore, MP, India
Jorge L. Hernández-Ardieta, University Carlos III of Madrid, Spain
Prof. Dr.C.Suresh Gnana Dhas, Anna University, India
Mrs Li Fang, Nanyang Technological University, Singapore
Prof. Pijush Biswas, RCC Institute of Information Technology, India
Dr. Siddhivinayak Kulkarni, University of Ballarat, Ballarat, Victoria, Australia
Dr. A. Arul Lawrence, Royal College of Engineering & Technology, India
Mr. Wongyos Keardsri, Chulalongkorn University, Bangkok, Thailand
Mr. Somesh Kumar Dewangan, CSVTU Bhilai (C.G.)/ Dimat Raipur, India
Mr. Hayder N. Jasem, University Putra Malaysia, Malaysia
Mr. A.V.Senthil Kumar, C. M. S. College of Science and Commerce, India
Mr. R. S. Karthik, C. M. S. College of Science and Commerce, India
Mr. P. Vasant, University Technology Petronas, Malaysia
Mr. Wong Kok Seng, Soongsil University, Seoul, South Korea
Mr. Praveen Ranjan Srivastava, BITS PILANI, India
Mr. Kong Sang Kelvin, Leong, The Hong Kong Polytechnic University, Hong Kong
Mr. Mohd Nazri Ismail, Universiti Kuala Lumpur, Malaysia
Dr. Rami J. Matarneh, Al-isra Private University, Amman, Jordan
Dr Ojesanmi Olusegun Ayodeji, Ajayi Crowther University, Oyo, Nigeria
Dr. Riktesh Srivastava, Skyline University, UAE
Dr. Oras F. Baker, UCSI University - Kuala Lumpur, Malaysia
Dr. Ahmed S. Ghiduk, Faculty of Science, Beni-Suef University, Egypt
and Department of Computer science, Taif University, Saudi Arabia
Mr. Tirthankar Gayen, IIT Kharagpur, India
Ms. Huei-Ru Tseng, National Chiao Tung University, Taiwan

CALL FOR PAPERS
International Journal of Computer Science and Information Security

IJCSIS 2013
ISSN: 1947-5500
http://sites.google.com/site/ijcsis/

International Journal Computer Science and Information Security, IJCSIS, is the premier scholarly venue in the areas of computer science and security issues. IJCSIS 2011 will provide a high profile, leading edge platform for researchers and engineers alike to publish state-of-the-art research in the respective fields of information technology and communication security. The journal will feature a diverse mixture of publication articles including core and applied computer science related topics.

Authors are solicited to contribute to the special issue by submitting articles that illustrate research results, projects, surveying works and industrial experiences that describe significant advances in the following areas, but are not limited to. Submissions may span a broad range of topics, e.g.:

Track A: Security

Access control, Anonymity, Audit and audit reduction & Authentication and authorization, Applied cryptography, Cryptanalysis, Digital Signatures, Biometric security, Boundary control devices, Certification and accreditation, Cross-layer design for security, Security & Network Management, Data and system integrity, Database security, Defensive information warfare, Denial of service protection, Intrusion Detection, Anti-malware, Distributed systems security, Electronic commerce, E-mail security, Spam, Phishing, E-mail fraud, Virus, worms, Trojan Protection, Grid security, Information hiding and watermarking & Information survivability, Insider threat protection, Integrity
Intellectual property protection, Internet/Intranet Security, Key management and key recovery, Language-based security, Mobile and wireless security, Mobile, Ad Hoc and Sensor Network Security, Monitoring and surveillance, Multimedia security ,Operating system security, Peer-to-peer security, Performance Evaluations of Protocols & Security Application, Privacy and data protection, Product evaluation criteria and compliance, Risk evaluation and security certification, Risk/vulnerability assessment, Security & Network Management, Security Models & protocols, Security threats & countermeasures (DDoS, MiM, Session Hijacking, Replay attack etc,), Trusted computing, Ubiquitous Computing Security, Virtualization security, VoIP security, Web 2.0 security, Submission Procedures, Active Defense Systems, Adaptive Defense Systems, Benchmark, Analysis and Evaluation of Security Systems, Distributed Access Control and Trust Management, Distributed Attack Systems and Mechanisms, Distributed Intrusion Detection/Prevention Systems, Denial-of-Service Attacks and Countermeasures, High Performance Security Systems, Identity Management and Authentication, Implementation, Deployment and Management of Security Systems, Intelligent Defense Systems, Internet and Network Forensics, Large-scale Attacks and Defense, RFID Security and Privacy, Security Architectures in Distributed Network Systems, Security for Critical Infrastructures, Security for P2P systems and Grid Systems, Security in E-Commerce, Security and Privacy in Wireless Networks, Secure Mobile Agents and Mobile Code, Security Protocols, Security Simulation and Tools, Security Theory and Tools, Standards and Assurance Methods, Trusted Computing, Viruses, Worms, and Other Malicious Code, World Wide Web Security, Novel and emerging secure architecture, Study of attack strategies, attack modeling, Case studies and analysis of actual attacks, Continuity of Operations during an attack, Key management, Trust management, Intrusion detection techniques, Intrusion response, alarm management, and correlation analysis, Study of tradeoffs between security and system performance, Intrusion tolerance systems, Secure protocols, Security in wireless networks (e.g. mesh networks, sensor networks, etc.), Cryptography and Secure Communications, Computer Forensics, Recovery and Healing, Security Visualization, Formal Methods in Security, Principles for Designing a Secure Computing System, Autonomic Security, Internet Security, Security in Health Care Systems, Security Solutions Using Reconfigurable Computing, Adaptive and Intelligent Defense Systems, Authentication and Access control, Denial of service attacks and countermeasures, Identity, Route and

Location Anonymity schemes, Intrusion detection and prevention techniques, Cryptography, encryption algorithms and Key management schemes, Secure routing schemes, Secure neighbor discovery and localization, Trust establishment and maintenance, Confidentiality and data integrity, Security architectures, deployments and solutions, Emerging threats to cloud-based services, Security model for new services, Cloud-aware web service security, Information hiding in Cloud Computing, Securing distributed data storage in cloud, Security, privacy and trust in mobile computing systems and applications, **Middleware security & Security features:** middleware software is an asset on
its own and has to be protected, interaction between security-specific and other middleware features, e.g., context-awareness, **Middleware-level security monitoring and measurement:** metrics and mechanisms for quantification and evaluation of security enforced by the middleware, **Security co-design:** trade-off and co-design between application-based and middleware-based security, **Policy-based management:** innovative support for policy-based definition and enforcement of security concerns, **Identification and authentication mechanisms:** Means to capture application specific constraints in defining and enforcing access control rules, **Middleware-oriented security patterns:** identification of patterns for sound, reusable security, **Security in aspect-based middleware:** mechanisms for isolating and enforcing security aspects, **Security in agent-based platforms:** protection for mobile code and platforms, Smart Devices: Biometrics, National ID cards, Embedded Systems Security and TPMs, RFID Systems Security, Smart Card Security, Pervasive Systems: Digital Rights Management (DRM) in pervasive environments, Intrusion Detection and Information Filtering, Localization Systems Security (Tracking of People and Goods), Mobile Commerce Security, Privacy Enhancing Technologies, Security Protocols (for Identification and Authentication, Confidentiality and Privacy, and Integrity), Ubiquitous Networks: Ad Hoc Networks Security, Delay-Tolerant Network Security, Domestic Network Security, Peer-to-Peer Networks Security, Security Issues in Mobile and Ubiquitous Networks, Security of GSM/GPRS/UMTS Systems, Sensor Networks Security, Vehicular Network Security, Wireless Communication Security: Bluetooth, NFC, WiFi, WiMAX, WiMedia, others

This Track will emphasize the design, implementation, management and applications of computer communications, networks and services. Topics of mostly theoretical nature are also welcome, provided there is clear practical potential in applying the results of such work.

Track B: Computer Science

Broadband wireless technologies: LTE, WiMAX, WiRAN, HSDPA, HSUPA, Resource allocation and interference management, Quality of service and scheduling methods, Capacity planning and dimensioning, Cross-layer design and Physical layer based issue, Interworking architecture and interoperability, Relay assisted and cooperative communications, Location and provisioning and mobility management, Call admission and flow/congestion control, Performance optimization, Channel capacity modeling and analysis, Middleware Issues: Event-based, publish/subscribe, and message-oriented middleware, Reconfigurable, adaptable, and reflective middleware approaches, Middleware solutions for reliability, fault tolerance, and quality-of-service, Scalability of middleware, Context-aware middleware, Autonomic and self-managing middleware, Evaluation techniques for middleware solutions, Formal methods and tools for designing, verifying, and evaluating, middleware, Software engineering techniques for middleware, Service oriented middleware, Agent-based middleware, Security middleware, Network Applications: Network-based automation, Cloud applications, Ubiquitous and pervasive applications, Collaborative applications, RFID and sensor network applications, Mobile applications, Smart home applications, Infrastructure monitoring and control applications, Remote health monitoring, GPS and location-based applications, Networked vehicles applications, Alert applications, Embeded Computer System, Advanced Control Systems, and Intelligent Control : Advanced control and measurement, computer and microprocessor-based control, signal processing, estimation and identification techniques, application specific IC's, nonlinear and adaptive control, optimal and robot control, intelligent control, evolutionary computing, and intelligent systems, instrumentation subject to critical conditions, automotive, marine and aero-space control and all other control applications, Intelligent Control System, Wiring/Wireless Sensor, Signal Control System. Sensors, Actuators and Systems Integration : Intelligent sensors and actuators, multisensor fusion, sensor array and multi-channel processing, micro/nano technology, microsensors and microactuators, instrumentation electronics, MEMS and system integration, wireless sensor, Network Sensor, Hybrid

Sensor, Distributed Sensor Networks. Signal and Image Processing : Digital signal processing theory, methods, DSP implementation, speech processing, image and multidimensional signal processing, Image analysis and processing, Image and Multimedia applications, Real-time multimedia signal processing, Computer vision, Emerging signal processing areas, Remote Sensing, Signal processing in education. Industrial Informatics: Industrial applications of neural networks, fuzzy algorithms, Neuro-Fuzzy application, bioInformatics, real-time computer control, real-time information systems, human-machine interfaces, CAD/CAM/CAT/CIM, virtual reality, industrial communications, flexible manufacturing systems, industrial automated process, Data Storage Management, Harddisk control, Supply Chain Management, Logistics applications, Power plant automation, Drives automation. Information Technology, Management of Information System : Management information systems, Information Management, Nursing information management, Information System, Information Technology and their application, Data retrieval, Data Base Management, Decision analysis methods, Information processing, Operations research, E-Business, E-Commerce, E-Government, Computer Business, Security and risk management, Medical imaging, Biotechnology, Bio-Medicine, Computer-based information systems in health care, Changing Access to Patient Information, Healthcare Management Information Technology. Communication/Computer Network, Transportation Application : On-board diagnostics, Active safety systems, Communication systems, Wireless technology, Communication application, Navigation and Guidance, Vision-based applications, Speech interface, Sensor fusion, Networking theory and technologies, Transportation information, Autonomous vehicle, Vehicle application of affective computing, Advance Computing technology and their application : Broadband and intelligent networks, Data Mining, Data fusion, Computational intelligence, Information and data security, Information indexing and retrieval, Information processing, Information systems and applications, Internet applications and performances, Knowledge based systems, Knowledge management, Software Engineering, Decision making, Mobile networks and services, Network management and services, Neural Network, Fuzzy logics, Neuro-Fuzzy, Expert approaches, Innovation Technology and Management : Innovation and product development, Emerging advances in business and its applications, Creativity in Internet management and retailing, B2B and B2C management, Electronic transceiver device for Retail Marketing Industries, Facilities planning and management, Innovative pervasive computing applications, Programming paradigms for pervasive systems, Software evolution and maintenance in pervasive systems, Middleware services and agent technologies, Adaptive, autonomic and context-aware computing, Mobile/Wireless computing systems and services in pervasive computing, Energy-efficient and green pervasive computing, Communication architectures for pervasive computing, Ad hoc networks for pervasive communications, Pervasive opportunistic communications and applications, Enabling technologies for pervasive systems (e.g., wireless BAN, PAN), Positioning and tracking technologies, Sensors and RFID in pervasive systems, Multimodal sensing and context for pervasive applications, Pervasive sensing, perception and semantic interpretation, Smart devices and intelligent environments, Trust, security and privacy issues in pervasive systems, User interfaces and interaction models, Virtual immersive communications, Wearable computers, Standards and interfaces for pervasive computing environments, Social and economic models for pervasive systems, Active and Programmable Networks, Ad Hoc & Sensor Network, Congestion and/or Flow Control, Content Distribution, Grid Networking, High-speed Network Architectures, Internet Services and Applications, Optical Networks, Mobile and Wireless Networks, Network Modeling and Simulation, Multicast, Multimedia Communications, Network Control and Management, Network Protocols, Network Performance, Network Measurement, Peer to Peer and Overlay Networks, Quality of Service and Quality of Experience, Ubiquitous Networks, Crosscutting Themes – Internet Technologies, Infrastructure, Services and Applications; Open Source Tools, Open Models and Architectures; Security, Privacy and Trust; Navigation Systems, Location Based Services; Social Networks and Online Communities; ICT Convergence, Digital Economy and Digital Divide, Neural Networks, Pattern Recognition, Computer Vision, Advanced Computing Architectures and New Programming Models, Visualization and Virtual Reality as Applied to Computational Science, Computer Architecture and Embedded Systems, Technology in Education, Theoretical Computer Science, Computing Ethics, Computing Practices & Applications

Authors are invited to submit papers through e-mail ijcsiseditor@gmail.com. Submissions must be original and should not have been published previously or be under consideration for publication while being evaluated by IJCSIS. Before submission authors should carefully read over the journal's Author Guidelines, which are located at http://sites.google.com/site/ijcsis/authors-notes .

ISSN 1947 5500
http://sites.google.com/site/ijcsis/